Detecting Journalistic Framing and Attitudes on News Reporting about Climate Change

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

Ye Jiang

Department of Computer Science

Faculty of Engineering

PhD Thesis

May 2020

Declarations

The candidate confirms that the work submitted is his own. Some parts of the work presented in this thesis have been published in the following articles:

Jiang, Y., Wang, Y., Song, X. and Maynard, D. (2020, January). Comparing Topic-Aware Neural Networks for Bias Detection of News. In Proceedings of 24th European Conference on Artificial Intelligence (ECAI 2020). International Joint Conferences on Artificial Intelligence (IJCAI).

Jiang, Y., Petrak, J., Song, X., Bontcheva, K., Maynard, D. (2019, June). Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 840-844).

Jiang, Y., Song, X., Harrison, J., Quegan, S., Maynard, D. (2017, September). Comparing Attitudes to Climate Change in the Media using sentiment analysis based on Latent Dirichlet Allocation. In Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism (pp. 25-30).

The candidate confirms that the above jointly-authored publications are primarily the work of the first author. The role of the other authors was mostly editorial and supervisory.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Abstract

News media play a key role in shaping the public understanding of scientific issues. Different news media frame stories in various ways to promote a particular viewpoint, and may exaggerate emotional aspects in order to gain readership. In order to understand how news articles are framed to advance certain agendas and what are the overall attitudes of newspaper (news attitudes) articles about particular issues relating to sentiment, methods for detecting news attitudes and journalistic framing are urgently needed. Due to the increasing amount of digitised news, these methods need to be automatic. Meanwhile, climate change is a contentious and emotionally charged issue, which leads to strong polarisation of opinions, making it an ideal topic for analysing different news attitudes and framing styles in the news articles.

Traditional journalistic methods to identify news attitudes and framing styles are typically carried out manually and therefore restricted to small case studies. Although machine learning approaches enable us to automatically detect such characteristics of news articles on a large scale, they typically require large amounts of annotated data, which is costly and time-consuming to produce. Also, unlike other types of text resource (e.g. tweets, reviews, etc) which are normally restricted, either formally or informally, to a certain limited document size, news articles are naturally more flexible in terms of the number of paragraphs used to convey one or more points of view. Learning models therefore must have the ability to adapt to a large range of document sizes.

To address these issues, this thesis first combines a topic model with opinion mining techniques to automatically identify the topics of each news article, and then applies lexicon-based sentiment analysis to the news articles which have a similar main topic. The proposed methods are fully unsupervised and therefore require minimum manual assessment. The experimental results indicate that different news publications have different angles and attitudes toward certain aspects of climate change issues. Secondly, this thesis develops deep learning methods for detecting journalistic framing styles, which may not only tell us whether a piece of news is positive or negative, but also enable us to investigate how the article is structured to promote a certain side of the issue. Two particular journalistic framing styles are intensively compared: hyperpartisan framing and tendentious framing. Taking advantage of the existing hyperpartisan news dataset from the SemEval 2019 task 4: Hyperpartisan News Detection, this thesis develops several deep learning architectures in a hierarchical framework to optimise the accuracy of the learning model. Taking account of the similarity between tendentious framing and hyperpartisan framing, it also establishes a relatively small tendentious climate change news corpus, and applies transfer learning, which adapts the model trained on general political partisanship to the detection of tendentiousness in the climate change news. This minimises the annotation cost, and also enhances the model accuracy. In order to optimise the model performance, contextualised word embeddings are applied on top of different neural encoders, and topic model distributions are combined with a hierarchical framework, which outperforms other baselines.

This thesis develops automatic methods for detecting news attitudes and journalistic framing styles in the news articles. The proposed method for detecting news attitudes could enable journalists, social scientists and news consumers to have a better understanding of how news publications' viewpoints are different in terms of emotional aspects. Also, the hyperpartisan framing and tendentious framing can be accurately detected by the proposed model and transfer learning technique. Furthermore, since transfer learning could potentially enhance the model performance, the methods can be extended to other types of frames that might be encountered on any issue of public concern, by fine-tuning on other media framing corpora.

Acknowledgements

I would like to express most profound thanks to my supervisor, Dr. Diana Maynard. Without her kindest encouragement, enormous support, and patient proofreading of all my works, I would never have made this thesis what it is today. It is thanks to her that I embarked on this PhD journey, and opened a brand new life-time experience.

My warmest thanks go to Prof. Jackie Harrison and Prof. Shuan Quegan, for their immense knowledge, guidance and constant feedback, which shaped this project. I acknowledge the financial support from the Grantham Centre for Sustainable Futures through research scholarship and training resources.

I also thank Dr. Xingyi Song, Johann Petrak, and other team members in the GATE NLP group, for their endless support and advice. It was a fantastic experience to work with them, especially when we won the Hyperpartisan News Detection competition!

Finally, my sincere thanks to my wife, for her patience and listening to my grumbles. Also, a special thanks to her for bringing me a big baby boy, which gave my PhD a different meaning! Last but not least, my thanks go to my family, without their support I would not have been able to achieve any of this.

Contents

1 Introduction 2

1.1 Motivation . . . 3

1.2 Challenges and Objectives . . . 4

1.3 Research Questions . . . 6

1.4 Contribution . . . 6

1.5 Thesis Structure . . . 8

2 Related Work 10

2.1 Polarised Attitudes to Climate Change in the News Media . . 10

2.2 Opinion Mining on Climate Change News . . . 13

2.2.1 Opinion Mining for News Articles . . . 15

2.2.2 Opinion Targets and Topic Modelling . . . 18

2.2.3 Machine Learning and Lexicon-based Opinion Mining . 21

2.2.4 Opinion Mining vs Stance Detection . . . 25

2.3 Journalistic Framing on Climate Change . . . 27

2.4 Tendentious Framing Detection . . . 32

2.4.1 Traditional Methods for Framing Detection . . . 32

2.4.2 Tendentious Framing and Hyperpartisan Framing . . . 34

2.4.3 Transfer Learning for Tendentious Framing Detection . 35

2.4.4 Deep Learning for Tendentious Framing Detection . . . 36

2.5 Summary . . . 41

3 Opinion Mining on Climate Change News 42

3.1 Problem Definition . . . 42

3.2 Methods . . . 43

3.2.1 Data Collection . . . 44

3.2.2 Pre-Processing . . . 45

3.2.3 LDA Model . . . 45

3.2.4 Applying SentiWordNet . . . 47

3.3 Results and Discussion 49
3.4 Limitation and Future Works 51
3.5 Summary 52

4 ELMo Sentence Representation Convolutional Network 53
4.1 Problem Definition 54
4.1.1 Data 54
4.1.2 Document Size 54
4.2 System Description 57
4.2.1 Deep Contextualized Word Representation: ELMo . . . 57
4.2.2 Convolutional Layers 58
4.2.3 Batch Normalization 58
4.2.4 Max Pooling Layer 60
4.2.5 Dense Layer 60
4.3 Experiments 61
4.3.1 Data 61
4.3.2 Preprocessing 61
4.3.3 Fine-tuning 62
4.3.4 Ensemble Learning 63
4.3.5 Evaluation 63
4.4 Results and Discussion 65
4.5 Limitations and Future Work 66
4.6 Summary 67

5 Hierarchical Document Representation 68
5.1 Problem Definition 68
5.2 Related Work 71
5.3 Methodology 72
5.3.1 Baseline Models 73
5.3.2 Hierarchical Models 74
5.3.3 Embedding Generation 77
5.4 Experiments 78
5.4.1 Dataset 78
5.4.2 Pre-processing 80
5.4.3 Evaluation 80
5.5 Results and Discussion 81
5.6 Summary 83

6 Hierarchical Topic-Aware Neural Networks 84

6.1 Problem Definition . . . 84

6.2 Related Work . . . 86

6.3 Method . . . 87

6.3.1 LDA Distributions . . . 88

6.3.2 Model Specifications . . . 88

6.4 Experiment . . . 94

6.4.1 Dataset . . . 94

6.4.2 Pre-processing . . . 95

6.4.3 Results and Discussion . . . 96

6.5 Conclusion . . . 97

7 Transfer Learning: From Hyperpartisan News to Tendentious News 99

7.1 Corpus Construction . . . 99

7.1.1 Gold Dataset . . . 100

7.1.2 Annotation Scheme . . . 101

7.2 Transfer Learning . . . 104

7.2.1 Method . . . 105

7.2.2 Result and Discussion . . . 106

7.3 Case Study: Sentiment Analysis in the Tendentious Corpus . . 107

7.4 Summary . . . 109

8 Conclusion and Future Work 110

8.1 Addressed Challenges . . . 111

8.2 Research Questions . . . 113

8.3 Limitations . . . 114

8.4 Future Work . . . 116

A Comparing Attitudes to Climate Change in the Media using Sentiment Analysis based on Latent Dirichlet Allocation 119

B Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network 126

C Comparing Topic-Aware Neural Networks for Bias Detection of News 132

D Hierarchical Contextual Document Representation for Detecting Hyperpartisan News Articles (Submitted) 141

List of Figures

2.1 LDA model 19

2.2 Convolutional Neural Network Architecture for Sentence Classification [Zhang and Wallace, 2015]. 39

40

3.1 Silhouette analysis for LDA model 46

3.2 Example of SentiWordNet 3.0 online graphical representation of first sense of the word 'faithful' as an adjective [AL-Sharuee et al., 2017]. 48

3.3 Attitudes of four news sources to the Copenhagen Summit in 2009. 50

4.1 System architecture, *F/B vector* denotes Forward/Backward hidden state from BiLSTM layers. 56

5.1 Baseline model structure 73

5.2 Hierarchical model structure 74

5.3 Document length distribution 79

5.4 Classification accuracy of the training models with ELMo embeddings 82

5.5 Classification accuracy of the training models with BERT embeddings 82

6.1 LDA based non-hierarchical models structure 89

6.2 Hierarchical model structure 92

6.3 Coherence scores in 500 topics 94

7.1 Annotation instructions 102

7.2 Sentiment Ratio in Tendentious and Non-Tendentious News between Four Publications. 108

8.1 Example of the importance of words and sentences in a news article. The blue and red color respectively denote the importance of sentence and word, the darker the more important. 117

List of Tables

2.1 Examples of news snippets from two different reporting styles. 35

2.2 List of Word Embeddings 38

3.1 Statistics per publisher. 44

3.2 Example of topic list in The Guardian 2007. Each topic has 10 keywords with its weight which reflect how important a keyword is to that topic. 47

3.3 Example of topic-document combination. For each article, it shows the topic ID numbers with their corresponding probabilities. 47

3.4 Topics in the year of 2009. 49

3.5 Example sentences with sentiment polarity detected in the four news sources in 2009. 50

4.1 Examples of Hyperpartisan news and Non-Hyperpartisan news in the *by-article* dataset. 55

4.2 System comparison (accuracy). 64

5.1 Statistics of dataset 79

5.2 Performance comparison between models. Best values are marked in **bold**, standard deviations in parentheses 81

6.1 Performance comparison between models. The best model accuracy is marked in **bold** 96

6.2 Performance comparison between combinations of topic distributions. TW denotes the topic-word distribution, DT denotes document-topic distribution. The best model accuracy is marked in **bold** 97

7.1 Key statistics per publisher before removing very large/small word counts articles. 100

7.2 Key statistics per publisher after removing articles with very large/small word counts. 100
7.3 Final numbers of news article in each category and publication. 104
7.4 System comparison, 'base' denotes the baseline models, 'In' denotes the In-domain pre-training, 'Cross' denotes the Cross-domain pre-training. 106

Chapter 1

Introduction

Climate change, also called global warming, refers to the increase in global temperature caused by rising greenhouse gas emissions [Weart, 2008]. Although the cause of climate change has been extensively discussed and attributed to human activities, much controversy still centres on the issue of climate change regarding its existence or effects, especially with respect to political, economic and scientific complexities [Rice et al., 2018]. The increasing evidence of man-made climate change, however, has not resulted in a proportional increase in news coverage [DiPeso, 2006]. As the former BBC journalist Kirby explained:

- "Alarming or not, climate change is becoming an increasingly hard subject to sell in much of the media...Editors are simply bored with what they think is an old story they have heard before." [Anderson, 2009]

Market-driven journalism [Daniel, 1995] transforms news readers into customers, news into products, and news circulation into markets, pressurising the institutional rules of news media [Štětka, 2013]. Such pressure also influences how news is packaged and presented, and how the selections of news affect the way in which news consumers understand the issues behind the story.

Although news media have a democratic function to inform the audience on various social realities [Opperhuizen et al., 2018], they have also been seen to play a crucial role in shaping the public's understanding about scientific issues [Friedman et al., 1986], and selected and packaged climate change news may potentially also cause polarised understandings of climate change.

With the growth of internet infrastructure and news media digitalisation, online newspaper articles are now being generated much faster, and their readership is becoming wider than traditional print media. People now tend to have much easier access to different news topics [Somaiya, 2014], and such access also results in an increasing polarisation in the beliefs about climate change.

For instance, although 97% of climate scientists agree that human-caused climate change is happening [Cook et al., 2016], a survey found that 30% of Americans refused to believe that climate change is taking place, while 40% refused to believe that it is human-caused [Anthony et al., 2018].

Issues related to climate change are typically framed in news reports to have a particular angle or slant, rather than just being presented as factual events [Beattie and Milojevich, 2017]. For instance, a recent piece of research [Vu et al., 2019] found that richer countries tend to frame climate change as a political issue, while poorer countries often frame it as an international issue that needs to be addressed globally.

Journalistic framing refers to giving emphasis to certain aspects and downplaying others when producing content [Gitlin, 2003, Tankard Jr, 2001]. It has been asserted that bias [Entman, 2007], cultural differences [Vasalou et al., 2010] and ideologies [Patterson and Donsbagh, 1996] all influence the construction of frames, shaping news messages. Given such differences, news media typically present diverse accounts of news stories, and news publications present different angles on certain issues. Thus, detecting journalistic framing in news articles could enable news consumers, social scientists and journalists to have a more detailed understanding of how climate change issues are framed in news stories.

1.1 Motivation

Since the controversial nature of climate change issues in the news media; potentially leads to strong polarisation of opinions, this thesis focuses on two related factors, namely the attitudes and journalistic framing of news reports.

1. **News Attitudes**: News coverage of climate change may potentially influence the social consensus around climate change. Dramatic climate change coverage is designed to capture the audience's attention by emphasising the political conflict and devastating impacts of climate change, and downplaying other explanatory factors. In this vein, it is important to document the nature of climate change news since polarised environments may strongly affect how audiences evaluate information. In order to understand whether the attitudes of the news publication presented in climate change news are polarised or balanced, this thesis first looks at the news attitudes toward certain climate change issues in news articles. This is to identify how news attitudes could potentially reveal how the

individual's perceived social consensus of climate change is influenced and changed by news reports.

2. **Journalistic Framing:** News audiences are more likely to choose a news article which is most aligned with their pre-existing ideologies and beliefs when they are presented with a diverse selection. Since the commercial pressure on media has increasingly dominated the institutional rules of news media, attracting attention from news audiences is of key importance for news publications to survive in markets where the competition for audiences and advertising revenue is of paramount importance. Consequently, understanding news framing is helpful as it not only tells us whether a piece of news is positive or negative about an issue, but also investigates how the article is structured to promote a certain side of the issue. These frames shape public attitudes on a variety of topics, and the topics of climate change among them. Understanding and detecting the journalistic framing in news media, and the influences on the diverse perspectives of news media towards climate change, would potentially lead to improved understanding of how news media plays a role in climate change mitigation. Large-scale user studies are urgently needed in order to better understand how news consumers' beliefs on climate change are affected by polarised and framed news articles.

1.2 Challenges and Objectives

The overall objective of this thesis is to *automatically detect news attitudes and journalistic framing in climate change news articles.* To achieve this, it first looks at the potential challenges this raises, and then discusses how each challenge can be addressed, as follows:

1. **Context Complexity**: Unlike other text resources (e.g. customer reviews, tweets, etc.), which normally have limited document size (i.e. limited number of sentences or words) and an explicit opinion target (i.e. product, event or issue), news articles are naturally more flexible in terms of the number of paragraphs used to convey one or more points of view. In order to investigate the overall attitudes of news articles toward certain climate change issues, we need to understand what the target of the attitude is in each case, i.e. which aspect of climate change the attitude is about. Traditional social sciences/journalistic studies methods

evaluating news opinions and identifying news topics are typically carried out manually, and are therefore limited to relatively small case studies. The objective of this thesis, however, is to apply machine learning and natural language processing (NLP) techniques on a large scale, looking at thousands of news articles and studying their different attitudes toward climate change issues.

2. **Annotation is Costly**: Supervised machine learning techniques require annotated data for training the models. However, it is impractical to manually identify and analyse those attitudes on a large scale since annotating news articles for the attitudes and journalistic framing are both time-consuming and costly. To address this, this thesis first implements unsupervised topic models to automatically identify the attitudes, and then applies opinion mining techniques to analyse the overall attitudes of news articles toward certain aspect of climate change. Meanwhile, the concept of transfer learning can be exploited to improve the model performance by training one model from one dataset and transferring its knowledge to another with a similar domain. This thesis also aims to minimise the annotation cost by creating a small journalistc framing dataset about tendentious climate change news, and then uses transfer learning to tackle the data insufficiency issue.

3. **Model Optimisation**: Deep learning models often perform exceptionally well when there is a huge amount of data, but this is not the case when the training processes suffer from data insufficiency issues, since the large number of parameters of learning models could result in over-fitting. Consequently, the architectures and configurations of learning models need to be carefully investigated since the journalistic framing datasets are relatively small for most of the existing deep learning models. To address this, in order to generate effective document representation for news articles, this thesis aims to explore different model structures for encoding the various sizes of newspaper articles and to investigate how different configurations would affect the ability of model generalisation in this thesis.

1.3 Research Questions

This thesis focuses on using techniques from NLP, machine learning and journalistic studies to automatically detect attitudes and journalistic framing styles from news articles in their reporting of climate change issues. To address the research challenges above, the research questions are listed as follow:

1. **RQ1**: *How to automatically analyse the overall attitude of newspaper articles towards a certain climate change issue?*
2. **RQ2**: *How to automatically and accurately detect journalistic framing from climate change news?*
3. **RQ3**: *What are the optimal structures and configurations of the learning models in detecting journalistic framing in the news articles?*

This thesis addresses each of these questions in the following chapters of this thesis: **RQ1** in Chapter 3, **RQ2** in Chapter 4 and in Chapter 7, **RQ3** in Chapter 5 and Chapter 6.

1.4 Contribution

The contributions of this thesis can be summarised as follows:

- An algorithm is developed that combines a topic model with opinion mining techniques to automatically identify the main topic of news articles, and implement an opinion mining technique to measure the attitudes expressed in articles which have a similar topic. Since newspaper articles typically involve multiple topics when reporting about climate change, this thesis implements a topic model to automatically identify the most relevant topic (i.e., the topic which has the highest probability). It assumes that news articles are similar if their topics are also similar, thus those articles which have similar or the same topics can be clustered. Finally, this thesis implements opinion mining on those articles to measure their attitudes towards certain climate change topics.
- The ELMo Sentence Representation Convolutional (ESRC) Network is developed to automatically identify journalistic framing in the news articles. Traditional neural networks use token sequences as input, but this implies either a high computational cost when a very large maximum sequence length is used to fully represent the longest articles, or

potentially a significant loss of information if the sequence length is restricted to a manageable number of initial tokens from the document. To address this, the ESRC takes account of the document structural hierarchical information between word and sentence, and between sentence and document. The ESRC also takes advantage of the state-of-the-art word embedding ELMo to generate sentence-level and document-level representations simultaneously. The method using ESRC was also the winning entry in the International Workshop of Semantic Evaluation 2019 (SemEval2019) Task 4: Hyperpartisan News Detection.

- A Topic-Aware Neural Network is developed, which incorporates the distributions generated from an LDA model with a hierarchical framework. The performance of several popular neural encoders, such as CNN, attention-RNN, Transformer, is compared extensively, with/without incorporating hierarchical frameworks and LDA distributions. The performance of the ELMo and BERT representations in document classification is also compared.

- A transfer learning method is introduced, that transfers the knowledge from hyperpartisan news articles to tendentious news articles concerning climate change. Taking advantage of the existence of a corpus of hyperpartisan news and the similarity between hyperpartisan and tendentious news, since annotating a large scale (tendentious) news corpus from scratch would be costly, this thesis implements a transfer learning approach to enhance the model performance by having a bigger training set, while also reducing the size of the tendentious news corpus required. Different pre-training methods are compared, and fine-tuning techniques are developed for the transfer learning task.

- A tendentious news corpus for climate change news articles is established. Two different annotation schemes are compared for newspaper article annotation; The relatively complex five-point scale is found to encourage annotators to select the tendentiousness confidently, compared with a three-point scale (i.e., True, False or Neutral). Since the concept of tendentiousness is more complicated than concepts like sentiment analysis, the five-point scale gives the annotators more flexibility in their scoring in order to express their lack of certainty, rather than making a 'hard' choice.

1.5 Thesis Structure

The thesis is structured as follows:

- **Chapter 2: Related Work** presents an overview of the relevant literature. This chapter starts with a broad view of how climate change attitudes are polarised in news media, and then introduces the concept of opinion mining in news articles. It also summarises the earlier works on newspaper opinion mining, and the potential challenges of opinion mining on news articles. Then, it presents the relation between opinion target and topic, and discusses the methods of automatic topic identification. Different machine learning and lexicon-based opinion mining approaches are summarised and compared. Next, the chapter introduces journalistic framing and how it may influence a reader's beliefs about climate change, and summarising the earlier works on detecting journalistic framing. Finally, the tendentious and hyperpartisan framing styles are introduced, showing how these styles can be automatically detected by implementing transfer learning and deep learning methods.

- **Chapter 3: Opinion Mining on Climate Change News** presents the algorithm for automatically identifying topics and measuring the attitudes of different news publications. It first introduces the topic model, Latent Dirichlet Allocation (LDA), and then describes how LDA is implemented in this work to identify the articles with similar topics. Then, a general sentiment lexicon, SentiWordNet, is presented. This is used to assign sentiment labels to each news article automatically. Finally, the results by implementing such combinations are discussed.

- **Chapter 4: ELMo Sentence Representation Convolutional Network** presents the method used for participation in the International Workshop on Semantic Evaluation 2019 (SemEval 2019). It starts with introducing the contextualized word embedding, ELMo, and explains how it was implemented in the model. Then, it presents the model structures, and discusses the effectiveness of batch normalisation in the model. It also discusses the characteristics of data, the pre-processing method, and approaches implemented in this task. Finally, it discusses the results in general.

- **Chapter 5: Hierarchical Document Representation** presents a hierarchical framework for encoding newspaper articles. It first introduces the model architectures, and discusses the earlier works on document classification. Then, it introduces baseline models and the hierarchical models. It also compares the effectiveness of two state-of-the-art word embeddings, ELMo and BERT, with/without incorporating a hierarchical framework. Finally, it discusses the experimental results and concludes that a hierarchical framework could significantly improve the performance of document representation especially for long documents.

- **Chapter 6: Hierarchical Topic-Aware Neural Network** introduces a model structure taking account of the topic distributions from LDA model and also the hierarchical document structural information. It first presents the traditional document representation, and the earlier works on generating document representations for classification. Then, it introduces the model structure which combines LDA distributions with word representation and sentence representation separately in the hierarchical framework. Finally, it discusses the experimental results and findings.

- **Chapter 7: Transfer Learning: From Hyperpartisan News to Tendentious News** presents the transfer learning techniques for detecting tendentious news. It first introduces the corpus construction by explaining the statistics of the corpus and the annotation scheme. Then, it presents the transfer learning approaches, and compares the effectiveness of two different pre-training strategies. Finally, the transfer learning methods and the tendentious corpus are evaluated.

- **Chapter 8: Conclusion and Future Work** summarises the overall findings and offers suggestions for future work.

Chapter 2

Related Work

This chapter starts with a broader view of how the news media covers climate change issues in Section 2.1, and discusses how climate change issues are polarised in the news media. In order to analyse attitudes toward climate change from different news publications, this chapter introduces the concept of opinion mining in news articles in Section 2.2. There it first discusses the challenges of opinion mining in news, and describes different levels of opinion mining. Then, it presents the relation between opinion target and topic, and shows how topics are identified automatically from large volumes of news articles. This chapter also discusses different opinion mining approaches, looking at both machine learning and lexicon-based methods. After that, it presents the idea of journalistic framing in news reports, and investigates how journalistic frames might affect news consumers' beliefs about climate change, in Section 2.3. Finally Section 2.4 reviews the literature of traditional framing detection and its drawbacks, introducing the tendentious and hyperpartisan framing styles, and showing how these styles can be automatically detected by implementing transfer learning and deep learning methods.

2.1 Polarised Attitudes to Climate Change in the News Media

Climate change is a global issue which can both affect and be affected by every individual, thereby creating a sense of global community. In this vein, the topic of climate change motivates many individuals to act in ways to encourage global public engagement, as witnessed recently by the momentous rise to fame of climate change activist Greta Thurnberg. Millions of students around the world have been inspired by her strikes, and Greta has received support from climate activists, scientists, world leaders and even the Pope, who advised her to "continue" her work. However, critics also accused her of being manipulated by left-wing green extremists, and suggested that actions to mitigate climate change are not as "black or white" as she claims[1].

On the other hand, although research has found that more than 97% of climate scientists agree that human-caused climate change is happening [Cook et al., 2016], there are also sceptics who do not believe climate change is real. For instance, one survey found that 30% of Americans refused to believe that climate change is happening, while 40% refused to believe that it is human-caused [Anthony et al., 2018]. Such scepticism also partially aligns with political polarisation. For instance, in the US, a study found that 92% of registered Democratic voters believed climate change is real, but only 51% of registered Republican voters did [Anthony et al., 2018]. Consequently, understanding the framing of climate change news coverage in climate change news coverage is an important step towards understanding why and how public opinion has become increasingly polarised, despite increasing scientific consensus on the reality and anthropogenic sources of climate change.

In order to address the issue about what causes the polarised social consensus in climate change, Sherif [1936], Asch [1955] have shown that perceived social consensus — the degree to which an individual believes others in their social group agree about an issue — has a strong influence on people's own beliefs. Such consensus affects an individual's ideology about climate change, especially when people are uncertain about what to think or how to behave [Goldberg et al., 2019]. For instance, if people incorrectly believe most other people do not believe global warming is happening, then they likely will not believe it is happening either, and will be much less likely to perceive global warming as a serious risk or support policy action to reduce it. Also, people are normally influenced by their own social circle. However, Mildenberger and Tingley [2019] investigated the extent to which people misperceived global warming beliefs in the United States and China, and found that US respondents could significantly increase their support for a global climate treaty when exposed to information about the proportion of Chinese people's pro-environment beliefs. The consensus of environment-supporting beliefs was changed in the US respondents as the Chinese pro-environment information affected their in-group consensus. This shows that even people from outside

[1] https://www.thecourier.co.uk/fp/opinion/readers-letters/1057870/readers-letters-greta-thunberg-has-been-manipulated-by-left-wing-green-extremists/

their social circle could at least partially influence their in-group consensus. Consequently, such information could be potentially manipulated to influence the social consensus by the information source, such as news media or social media.

News coverage is likely to have a strong influence on public attitudes about climate change because traditional news reporting still remains the dominant way that the public learns about scientific issues [NSB, 2016]. For example, individuals typically link extreme weather to climate change because of the news reports [Durfee and Corbett, 2005]. Nevertheless, dramatic climate change coverage captures audiences' attention by focusing on "*embittered conflict, momentous events, or devastating impacts*", rather than persistent problems [Bennett et al., 2008]. For instance, a study of U.S. media found that scientific voices have become less prominent in climate change stories, and to a large extent have been replaced by politicians who are speaking more about policy options than science [Trumbo, 1996]. It seems that readers are more receptive to controversial claims about climate change when they are presented within a more complex context, rather than "just the facts" narratives [Corbett and Durfee, 2004].

Additionally, this tendency toward conflict politicises climate coverage by prominently featuring political actors as official sources that can speak for competing factions [Bennett et al., 2008], and as a result encourages individuals to follow political elites' opinions rather than those of scientists [Slothuus and De Vreese, 2010, Bolsen et al., 2014]. The increasingly politicized coverage of environmental issues [Boykoff, 2011, Boykoff and Luedecke, 2016] also results in polarisation of views about climate change, especially when political actors stand for different partisan viewpoints. For instance, the U.S. President, Trump, made public statements both formally and informally denying climate change is happening[2], and claimed that "the concept of global warming was created by Chinese in order to make US manufacturing non-competitive"[3]. In 2019, a survey[4] found that a total of 52% of Americans who described themselves as "very rightwing" believed that global warming was a hoax.

The polarisation in climate change news is important to document because polarised environments strongly affect how individuals evaluate information. Specifically, polarisation intensifies the impact of partisan elites on individuals' attitudes, while decreasing the impact of other substantive information,

[2] https://www.bbc.co.uk/news/world-us-canada-46351940

[3] https://twitter.com/realDonaldTrump/status/265895292191248385

[4] https://yougov.co.uk/topics/yougov-cambridge/globalism-project

leading individuals to become more confident in their less substantiated beliefs [Druckman et al., 2013]. This attitude towards polarisation has led to researchers paying increasing attention to the role that news coverage has played in shaping public opinion, particularly the ways in which polarisation in news coverage have affected public attitudes [Bolsen et al., 2014, Druckman et al., 2013].

In summary, this section has presented the linkage between social consensus of climate change and news coverage. Studies of the increasing politicisation and polarisation of climate change issues reveal that an individual's belief about climate change is affected by the in-group consensus, and that news media plays an important role in influencing and changing the perceived social consensus. In order to investigate how news media could affect public attitudes toward climate change, this thesis focuses on the different attitudes in news reports and how news media stories about climate change differ.

2.2 Opinion Mining on Climate Change News

To understand better the attitudes presented in climate change news, and whether they are polarised or balanced, this study first looks at the sentiment orientations in newspaper articles.

Sentiment orientation is a measure of subjectivity and opinion in text, which typically captures a sentiment factor (e.g., either positive or negative) and its intensity (i.e., the degree of how positive or negative the word, sentence or document is) towards a subject topic, person or idea [Osgood et al., 1957]. The analysis and automatic extraction of sentiment orientation can be defined as opinion mining [Taboada et al., 2011].

Opinion mining, also called sentiment analysis, is a field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organisations, individuals, issues, events, topics, and their attributes [Liu, 2012].

Opinions are one of the most essential elements to many human activities, and especially recently, the opinionated postings in social media have reshaped traditional business activities, as well as our political systems. Applications of opinion mining have gone beyond just looking at whether sentiment is positive or negative, and have spread to different domains, such as customer reviews, polls, and general elections. For instance, McGlohon et al. [2010] applied the results of sentiment analysis of customer reviews in order to rank products and merchants. Their data included 8 million product reviews and 1.5 million merchant reviews, where each review comes with a 5-scale rating which indicates the sentiment orientation. They first averaged the review ratings as a baseline model, and proposed re-weighting models that filtered reviews out or downgraded their influence in the composite score, considering some reviews to be more important or reliable than others. O'Connor et al. [2010] also linked Twitter sentiment with public opinion polls by collecting 1 billion tweets over the years 2008 and 2009. They used tweets that contained certain topic keywords for each poll, and counted instances of positive-sentiment and negative-sentiment words in the context of a topic keyword. Tumasjan et al. [2010] also used Twitter sentiment to predict election results. They used a lexicon-based tool, Linguistic Inquiry and Word Count (LIWC 2007), to investigate over 100,000 messages containing a reference to either a political party or a politician.

Several sources of data have been analysed for opinion mining, and in particular social media, such as Twitter, has been heavily used for opinion mining research by taking advantage of the large volume of data and its easy accessibility. On the one hand, recent statistics[5] indicates that almost 3.8 billion online users are using social networking sites (e.g., Facebook, tweets, etc.) to share information, thoughts and opinions in 2020. With the growing amount of such data, it provides rich and useful information for research in social activities. For instance, a study invented a Germtracker [Sadilek et al., 2012] which derives accurate real-time epidemiological information from tweets to predict who might get flu, and which restaurants might have high risk of food poisoning. On the other hand, because tweets contain a short piece of text which is limited to 280 characters, the majority of them are therefore considered as conveying a single topic due to this length limitation [Giachanou and Crestani, 2016b,a], and the complexity of their context thereby could be simplified as only one opinion per tweet. For instance, Kouloumpis et al. [2011] collected 2000 tweets for movie reviews, and considered each tweet as either positive or negative with respect to only one movie. Wang et al. [2012] manually constructed rules that were simple logical keyword combinations to retrieve relevant political tweets (e.g., tweets for Mitt Romney) and assigned a single sentiment to each tweet.

[5] https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

2.2.1 Opinion Mining for News Articles

Newspaper articles, however, are naturally more flexible in terms of the number of paragraphs used to convey one or more points of view [Jiang et al., 2020]. Specifically, conducting opinion mining on newspaper articles is difficult because the content of news reports is much more complicated, due to their diverse context length (i.e., the huge variation in word counts in news articles), and because the topics of the news are normally varied (i.e., talking about different aspects of an event), compared with other types of sources [Jiang et al., 2017]. Thus, the polarities of some news articles towards the given topic could be misclassified because of the interference from sentiments towards other topics in the articles.

Research on opinion mining for news articles is not entirely new. However, taking account of the complexity of news content and its topical diversity, instead of considering the entire article, much research has only focused on the small snippets of news articles, such as news titles [Burscher et al., 2016, Agarwal et al., 2016], news comments [Rahab et al., 2017, Pérez-Granados et al., 2012, Maynard et al., 2014], or quotations [Balahur et al., 2009, 2013]. For instance, Agarwal et al. [2016] implemented an off-the-shelf tool, SentiWordNet 3.0, on around 500 news headlines and classified these headlines into a binary sentiment orientation. Burscher et al. [2016] combined a clustering algorithm with sentiment analysis to analyse sentiment orientation in the news headlines. They collected 4,286 news articles, and clustered each news article into several news frames based on the features extracted from the headlines. Then they implemented SentiWordNet to assign sentiment scores for each article in the clusters. Comments posted on the online news pages have also been discussed extensively. Rahab et al. [2017] collected 147 comments from Arabic newspapers on the web and manually annotated a binary class to each comment, then trained machine learning classifiers to classify these comments. Also, Pérez-Granados et al. [2012] crawled news comments based on pre-defined rules and automatically assigned sentiment labels for the comments based on pre-trained classifiers. Text in quotes is usually more subjective than the other parts of news articles as they are often referred to as the source of the opinion statement. Balahur et al. [2009] collected a set of 99 quotes based on the agreement between a minimum of two annotators, and annotated each quote into binary classes, then applied a machine learning classifier on these quotes. In their later experiments, Balahur et al. [2013] expanded the data set to 1592 news quotes, and labelled each quote with respect to binary classes, similar to their previous work.

Since news can refer to multiple topics and, consequently, have multiple polarities, de Arruda et al. [2015] split news into smaller units of annotation to capture each of these individual topics. This, however, is still an unsettled issue, with current approaches typically segmenting news into sentences and presumably then finding the sentiment for each sentence separately [Balahur et al., 2013, Abdul-Mageed and Diab, 2012]. However, snippets cannot always reveal the insights of news articles, missing much useful information [Jiang et al., 2017] especially in news reporting about climate change. For instance, research has found that news headlines related to climate change typically depict fear, misery and doom [Boykoff, 2008], and describe climate change as sensational, alarming and harmful [Carvalho and Burgess, 2005, Zamith et al., 2013]:

- "*Final call to save the world from 'climate catastrophe'.*" — BBC, 8 Oct 2018.
- "*How much longer do we have left on Earth*". — Independent, 18 Apr 2020.
- "*Probably the worst year in a century: Australia's environmental toll of 2019*". — Guardian, 29 Mar 2020.

Sensational headlines, which evoke emotion or inspire curiosity depending on the subject, are frequently used to attract attention from the audience. However, the narratives in the sensational headlines are normally also subtle and implicit in order to retain curiosity to its readers, and therefore contain much fewer explicit entities or aspects.

Generally, opinion mining on newspaper articles can be divided into three levels:

- **Document level**: It aims to classify whether the whole opinion of a document expresses either a positive or negative sentiment [Pang et al., 2002, Turney, 2002]. This task is typically known as *document-level sentiment classification*, and assumes that each document expresses opinions on a single entity. For instance, given a movie review, the system determines whether the review expresses an overall positive or negative opinion about the movie. Commonly, newspaper articles use document-level opinion mining [Burscher et al., 2016, Balahur et al., 2009], where the overall sentiment orientation of a news article is computed based on the annotated sentiment orientation of each word in the document [Jiang et al., 2017]. Unlike other types of text resources, such as tweets or customer reviews which typically have a concrete object (i.e., a single opinion target), news articles may span larger subject domains, more complex event descriptions, and a whole range of topics [Balahur et al., 2013]. This is problematic when analysing the sentiment of a news article that includes many topics, as it is difficult to identify which topic contributes what sentiment to the whole article.

- **Sentence level**: The sentence-level sentiment classification aims to determine whether each sentence expresses a positive, negative, or neutral opinion [Liu, 2012]. This is similar to document-level opinion mining, but focuses more on a short piece of information, such as tweets [Pak and Paroubek, 2010, Earle et al., 2012, Bollen et al., 2011], reviews [McGlohon et al., 2010] and news headlines [Burscher et al., 2016, Agarwal et al., 2016], etc. Sentence-level sentiment classification is normally considered where there is a single source of opinion per sentence, and the overall sentiment orientation of a sentence is computed based on the annotated sentiment orientation of each word in the sentence.

- **Aspect level**: This is a fine-grained opinion mining task that aims to discover what exactly people liked and did not like. Unlike document-level or sentence-level which looks at the language constructs (e.g., documents, paragraphs, sentences and phrases), this looks at the opinion itself and assumes that an opinion consists of a sentiment (positive or negative) and a target (of the opinion). An opinion without its target being identified is of limited use. For instance, although the sentence *"Climate change is threatening wildlife, but the warming temperature is also reducing death rate for some animals in winter."* clearly has a negative tone, we cannot say that this sentence is entirely negative as the sentence is negative about *"climate change"*, but positive about *"warming temperature"*. Thus, the goal of aspect-level opinion mining is to discover sentiments on the opinion target, which are described by entities and/or their different aspects. This level of analysis has also been extensively conducted on Twitter [Salas-Zárate et al., 2017], reviews [Singh et al., 2013] and in different languages [Steinberger et al., 2014].

Since news articles have multiple entities/topics, and consequently, have multiple opinion targets, thus opinion summarisation [Maynard et al., 2014] needs to be computed by either, for instance, aggregating all the opinions of each sentence in the news article, or calculating the opinions based on each sentence/paragraph that contains aspects/entities in the article. The former, however, still ignores the opinion targets and leads to the overall opinion of the news article remaining unclear. The latter requires extra work on identifying the opinion targets before conducting opinion mining.

This thesis implements a hybrid method that takes advantage of both document-level and aspect-level opinion mining, and which calculates the overall sentiment orientation of a news article by taking account of the main opinion target of each news article (see Chapter 3). It assumes that the overall opinion target is equivalent to the main topic of the news article, so our primary goal is to automatically identify the main topic, which has the highest probabilistic distribution from the topic model.

2.2.2 Opinion Targets and Topic Modelling

Traditionally, document-level opinion mining in news articles either assume each article has a broad and fixed opinion target, or use machine learning approaches to extract opinion targets automatically. An example of the former is Im et al. [2013], who collected newspaper articles with their topic restricted to financial markets (using keywords such as "*earning announcements*" or "*quarterly reports*") from Malaysian local online news. Some other news corpora (e.g., 20newsgroup [Lang, 1995], SIAAC [Rahab et al., 2017]) were directly generated based on some broad or general topics, such as sport, medicine, science, etc. Such methods are typically based on the keywords of the news article,but did not consider how the keywords correlate to the article (i.e., the frequency of the word in the article). To address this, machine learning approaches have been implemented for extracting the keywords of the news article. Burscher et al. [2016] implemented opinion mining on nuclear power debate-related newspaper articles, and performed unsupervised k-means clustering to automatically cluster articles based on similar words in each article. Although news articles could be categorised by calculating the Euclidean distance of each keyword in the vector space (e.g., using term frequency-inverse document frequency (TF-IDF)), the importance of each keyword to the article (i.e., the probabilistic distribution of each word in an article) is still missing.

Topic modelling is a technique for discovering the most frequent words in a collection of documents by using statistical models. Probabilistic topic models, such as Latent Dirichlet Allocation (LDA) [Blei et al., 2003], have been used in the field of information retrieval to analyse the content of documents and the meaning of words. Traditionally, topic extraction requires reading a large number of articles manually, and it is not feasible to be implemented on a large scale (i.e., thousands of news articles). In the field of Natural Language Processing (NLP), LDA allows us to automatically classify documents and estimate their relevance to various topics. More importantly, LDA can generate topics from a large volume of text with minimal human assessment. For instance, Newman et al. [2006] implemented an LDA model to extract topics from 330,000 news articles, and analysed these topics, topic trends, and topics that relate entities by combining them with named entity recognisers. Feuerriegel et al. [2016] applied LDA to analyse how topics in financial news affect stock prices. They collected 7,645 news announcements and extracted 40 topics for each of them. Lin and He [2009] introduced a Joint Sentiment-Topic (JST) model that assumed each sentiment has a multinomial distribution over topics and also has multinomial distribution over words; They also extended the JST model by incorporating sentiment prior knowledge based on pre-compiled sentiment lexicons [Lin et al., 2011].

Specifically, LDA is a generative probabilistic topic model for collections of discrete data such as text corpora. In the LDA model, each document is a mixture of topics, and each topic is a probability distribution over words. LDA is a unsupervised method and has previously been implemented for topic identification in newspaper articles [Llewellyn et al., 2014, DiMaggio et al., 2013, Maier et al., 2018]. Figure 2.1 demonstrates the generative process in the LDA model.

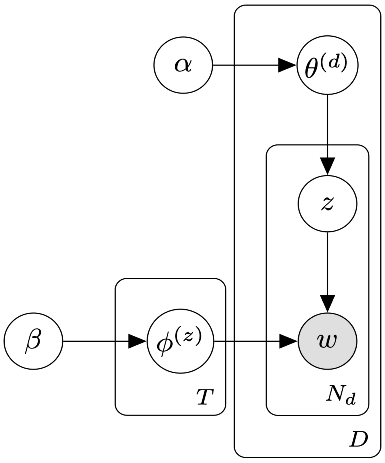


Figure 2.1: LDA model

Given a corpus C with a collection of D documents, which can be denoted by $C = \{d_1, d_2, ..., d_D\}$, each document in the corpus is a sequence of N_d words, thus a document d is denoted by $d = \{w_1, w_2, ..., w_{N_d}\}$, and each word in the document is an item from a vocabulary with total number of words V denoted by $\{1, 2, ..., V\}$. Finally, let T be the total number of topics. The procedure for generating a word in a document thereby can be broken down into two steps: a) choose a distribution over a mixture of T topics for the document; b) pick a topic randomly from the topic distribution, and draw a word from that topic according to the corresponding topic-word distribution. The formal definitions can be described as follow:

- For each topic $j \in \{1, 2, ..., T\}$, draw a distribution $\phi_j \sim Dir(\beta)$
- For each document $d \in \{1, 2, ..., D\}$, draw a distribution $\theta_d \sim Dir(\alpha)$
- For each word w_i in document d, draw a topic $z_i \sim Mult(\theta_d)$ and draw a word $w_i \sim Mult(\phi_{z_i})$

The observed variables are just the words in each document d; the others are latent variables (i.e., θ and ϕ) and hyperparameters (i.e., α and β) in the LDA model. LDA aims to infer or compute the distributions of those latent variables on the documents. Given a generative model, a number of topics, and some data, the process of uncovering the hidden variables of the LDA model is called inference. Formally, the target inference is the posterior distribution $P(z|w)$ (i.e., the probability of a topic assignment z given a corpus w). Many algorithms have been developed for posterior inference, such as message passing [Zeng et al., 2012], variational inference [Blei et al., 2003], gradient descent [Hoffman et al., 2010] and Gibbs sampling [Griffiths and Steyvers, 2004]. Among the above, Gibbs sampling is the most widely used inference method with LDA, because of its simplicity and fast optimisations specific to topic models [Yao et al., 2009].

This thesis implements LDA for two objectives:

1. **Topic Generation for Opinion Mining**: Climate change is a broad topic which typically contains many subtopics (e.g., pollution, carbon emission, etc). The contents of news articles therefore typically refer to different kinds of subtopics when talking about climate change. In order to identify various aspects of the climate change issue (i.e., opinion targets) mentioned in the newspaper articles and the attitudes toward such issues, this study implements LDA for automatic topic extraction on a large volume of news articles. Then, it regroups the generated topics from LDA based on their similarities, and conducts opinion mining on the articles which have similar topics (see Chapter 3).

2. **Feature Enrichment for Deep Learning**: There are two statistical distributions that are generated from LDA: 1) the document-topic distributions $\theta = \{\theta_d\}_{d=1}^{D}$ and 2) the topic-word distributions $\phi = \{\phi_j\}_{j=1}^{T}$. Normally, the topic-word distributions can be used to identify the top N words for each topic, and the document-topic distribution can be used to classify a document based on the topic having the highest proportion in each document. Such characteristics of those distributions could be provided as additive features for deep learning models and therefore enhance the model performance. This thesis proposes a hierarchical topic-aware model which takes account of those distributions on both the word level and sentence level independently, and the results outperform others in news article framing detection (see details in Chapter 6).

To conclude, the topics generated from the LDA model allow us to understand what are the opinion targets in the newspaper article, and furthermore, provide insight about which topic is more likely in each article and what are the keywords in the topics. Meanwhile, LDA is a fully unsupervised method, which require minimal human assessment, so that it can reduce annotation cost. After obtaining the main topic from each article, the method starts with opinion mining on each article to investigates what are the attitudes from different news publications toward the climate change issue.

2.2.3 Machine Learning and Lexicon-based Opinion Mining

Machine learning methods for opinion mining can be categorised into three groups: supervised [Pang et al., 2002, Pang and Lee, 2005], semi-supervised [Ortigosa-Hernández et al., 2012], and unsupervised [Turney, 2002].

- **Supervised** learning aims to build a classifier which can automatically learn the feature patterns from a labelled dataset. It normally needs feature engineering, which transfers text into a numerical representation and aims to remove the irrelevant or redundant features from the dataset in order to improve the performance of the classifier, to build a feature representation such as N-grams, TF-IDF or Bag-of-Words (BOW) for training the classifiers. Some of the commonly used classifiers for supervised learning are Support Vector Machine (SVM), Naive Bayes (NB), Neural Networks (NN) and Maximum Entropy (ME). For instance, Pang et al. [2002] pioneered using a machine learning classifier, such as NB, ME and SVM, on movie review sentiment analysis. They collected 752 negative and 1301 positive reviews from the IMDb archive, and randomly selected 700 positive reviews and 700 negative reviews. They divided data into three equal-sized folds and maintained a balanced class distribution in each fold. Their experiments were tested with various feature engineering options, where SVM with unigram features yielded the highest accuracy of 82.9%. Similarly, Zhang et al. [2011] used SVM and NB classifiers with different feature representations like unigram, bigram and trigram, etc, for sentiment classification in restaurant reviews. Although supervised opinion mining often achieves state-of-the-art performances, it relies on the availability of a labelled training data set, and thus typically requires extra human efforts to do this annotation.

- **Semi-supervised** opinion mining addresses the data insufficiency issue by using a large amount of unlabelled data together with a small amount of labelled data to build a training model. It aims to label unlabelled data using knowledge learned from that small amount of labelled data. For instance, a recent study Miao et al. [2020] implemented a semi-supervised method that trains the model with 3045 labelled sentences and 30,450 unlabelled sentences on Aspect-Based Sentiment Analysis datasets. Similarly, Yu and Kübler [2011] built a semi-supervised model on IMDB movie reviews, the Wall Street Journal and the JDPA Blog corpus. They randomly split each dataset into 90% unlabelled data, i% ($1 \leq i \leq 5$) labelled data, and 5% test data. Consequently, such methods are normally used when labelling the data or gathering labelled data is too difficult or expensive. However, such a method is not always "*the hammer to the nail*" solution, and the unlabelled data needs to have a strong correlation with the labelled data. For instance, research [Van Engelen and Hoos, 2020] has claimed that unlabelled data might also lead to worse performance in the classifier if wrong assumptions are made, and Singh et al. [2009] has claimed that the unlabelled data can help learning only in certain situations (i.e., when the margin between separated sets is large enough compared with average spacing between unlabelled data points), while in other situations they may not help.

- **Unsupervised** learning is a type of machine learning algorithm used to draw inferences from data sets consisting of input data without labelled responses. It aims to infer the natural structure present within a dataset. Some of the commonly used methods for unsupervised learning are clustering and association rule mining. Clustering aims to discover similar items based on the features of each cluster. For instance, Unnisa et al. [2016] developed a spectral clustering method to classify tweets into positive or negative clusters based on the word distance in the vector space model. They collected 2000 tweets about movie reviews, and selected the presence of each character and negation words as features for creating feature vectors. For association rule mining, the method aims to discover the relations between variables in a large unstructured dataset. For instance, Kim et al. [2009] extracted four types of association rules to indicate the opinion of customers for products and aspects of products that appear frequently in product reviews. They calculated the sentiment orientation by using the Pointwise Mutual Information (PMI) [Church and Hanks, 1990] algorithm on the rules that have a high confidence value. Although unsupervised methods do not require labelled data, the interpretation of their results is sometimes difficult since no gold-standard answer is available. For instance, clustering analysis does not have a solid evaluation metric (e.g., accuracy, F-measures, etc) as is used in supervised learning, as there is no ground truth (i.e., labels). Some clustering methods, such as K-means, require one to manually define the number of cluster K before the algorithm is applied. The exact number of clusters might affect its performance, and sometimes this needs domain knowledge and intuition [Pham et al., 2005].

Lexicon-based methods use opinion words, which are words that are commonly used to express positive or negative opinions (e.g., good, bad, amazing, etc.), in order to perform opinion mining [Hu and Liu, 2004]. Specifically, lexicon-based methods normally count the number of opinion words that are near the opinion targets in the text. For instance, if there are more positive opinion words than negative opinion words, the final opinion for the opinion target is positive, and otherwise negative [Ding et al., 2008]. The set of opinion words, constituting an opinion lexicon, is normally obtained by a bootstrapping method. For instance, Banea et al. [2008] built subjectivity lexicons by starting with a set of seed words, and expanding the seed set with related words found in an online dictionary. They then filtered this using a measure of word similarity, and continued to the next iteration until a maximum number of iterations was reached.

Traditionally, there are two types of lexicon-based approach for opinion mining on the sentence-level and document-level:

1. **Corpus-based approaches**: These aim to find co-occurrence patterns of words to determine the sentiment orientation of words or phrases in a sentence or document, and assume that the information in the context surrounding the opinion target might be exploited to help with polarity assignment [Shelke et al., 2012]. For instance, Mihalcea and Liu [2006] implemented a corpus-based method to assign a '*happiness*' factor to words based on the frequency of their occurrences in '*happy*' labelled blog posts compared with their total frequency in the corpus which contains both '*happy*' and '*sad*' labelled blog posts.

2. **Dictionary-based approaches**: These approaches normally utilise lexical resources, such as WordNet, to extract opinion words automatically. They use synonyms and antonyms in the lexical resources to determine the sentiment orientations of a word based on a set of opinion words [Shelke et al., 2012]. For instance, Mihalcea and Strapparava [2005] used WordNet-Affect, an extension of WordNet, to detect emotion in text automatically. They used weights from WordNet-Affect to assign sentiment weights to directly affective words, and assign affective weights to other words based on their similarity to an emotional category. There are several popular resources for dictionary-based approaches :

    - **SentiWordNet**: This is a specific lexical resource devised to support sentiment analysis applications, and also an extension of WordNet. The annotation of opinion words is based on three numerical scores (positivity, negativity and neutrality) for each WordNet synset, and different senses of the same word may have different sentiment scores.
    - **WordNet-Affect**: This is also an extension of WordNet, which labels affective-related synsets with affective concepts. For instance, the term euphoria is labelled with the concept '*positive-emotion*', and the noun illness is labelled with physical state, etc.
    - **MPQA**: This consists of 8,222 terms, which are labelled as subjective expressions, obtained from several sources. It contains a list of words, with their Part-Of-Speech tag, polarity (positive, negative, neutral) and intensity (strong, weak).
    - **SenticNet**: This was designed for aspect-level opinion mining based on the Sentic Computing [Cambria and Hussain, 2012] paradigm.

> SenticNet consists of 14,000 common sense concepts with a sentiment score in a range between -1 and 1. It is not only able to associate polarity and affective information, but also to deal with complex concepts, such as '*accomplishing goal*', '*celebrate special occasion*', etc.

Compared with the dictionary-based approach, the corpus-based approach has the advantages of being able to generate a domain-specific lexicon [Alqasemi et al., 2018], and is able to capture informal terms and slang [Peng and Park, 2011] as is typically found in social media. However, the corpus-based approach typically requires a massive corpus to capture the entire span of vocabulary words across different domains, and thus might result in being computationally intensive. For example, Turney and Littman [2003] used a 100 billion word corpus to obtain good accuracy from the PMI algorithm. Also, the co-occurrence statistics might not always be reliable. Kanayama and Nasukawa [2006] claimed that only about 60% of co-occurrences reflect similar sentiment, and antonyms of adjectives often co-occur together in the same phrase or sentence [Gross and Miller, 1990].

In summary, this section has presented the different types of methods for opinion mining and investigated their pros and cons. Based on this, the approach in this thesis therefore uses lexicon-based opinion mining on climate change-related news articles for a number of reasons. First, although machine learning classifiers perform quite well in the domain that they are trained on, the performance typically drops precipitously when the same classifier is used in a different domain [Aue and Gamon, 2005, Pang et al., 2008]. As discussed above, newspaper articles, unlike other sources (e.g., tweets, reviews), typically have a number of subtopics, and different domains might even be discussed in a single article. Second, machine learning classifiers cannot accurately tell the differences in the local context of a word, such as negation (e.g., *'not good'*) and intensification (e.g., *'very good'*) [Kennedy and Inkpen, 2006]. Third, machine learning classifiers typically need labelled data for training a classifier; however, the annotation is both time-consuming and costly, and there is no such type of data available for this study.

2.2.4 Opinion Mining vs Stance Detection

Stance detection aims to identify the expression of the individual's viewpoint toward an certain event or entity [Biber and Finegan, 1988], for instance, to detect whether the stance is supporting or against a topic, while opinion mining refers more to investigating the emotional state of the individual that expressed the opinion [Pang et al., 2008]. Although most of the studies in stance detection have focused on using the emotional state of the text to infer the stance [Ebrahimi et al., 2016, Elfardy and Diab, 2016, Lee, 2018, Overbey et al., 2017, Tsolmon et al., 2012, Unankard et al., 2014], another line of research develops a stance detection approach where sentiment is omitted [Darwish et al., 2017, Trabelsi and Zaiane, 2018, Darwish et al., 2020]. They found that using sentiment as a sole factor for the stance detection might be suboptimal, and even indicate a rather weak relation between sentiment and stance [Mohammad et al., 2017, Elfardy and Diab, 2016].

Many studies have used opinion expressed in a given text to indicate the stance interchangeably [Trabelsi and Zaiane, 2018, Unankard et al., 2014]. The sentiment polarity has been used to detect the stance towards various events in social media. For instance, Smith et al. [2017] annotated the sentiment expressed in the tweet to identify the opinion towards the terrorist attack in Paris in 2015. They labelled the tweets as negative, positive or neutral and analyse the public reaction to the attack. Park et al. [2011] used the sentiment to identify the political leaning of the commenter on news articles. They constructed a sentiment profile of each commenter to track their polarity towards a political party, and found that liberal commenters typically expressed negative viewpoints in conservative news articles but positive viewpoints in liberal articles. A recent study [Lee, 2018] categorized 25 most common hashtags with sentiment polarity about 2016 US presidential election day, and another [Agarwal et al., 2018] used the AFINN-111 dictionary to analyse the sentiment and used sentiment polarity as an indication of the stance towards Brexit. The sentiment has thus been seen as the indicator of the stance toward the event or topic in the above studies.

However, another line of studies used sentiment as proxy to help identify the stance [Ebrahimi et al., 2016, Elfardy and Diab, 2016, Mohammad et al., 2017]. For instance, the SemEval stance detection [Mohammad et al., 2016] labelled the tweets with sentiment and stance separately, and found that sentiment is a useful feature for stance classification when it is combined with other features and not used alone. Igarashi et al. [2016] used SentiWordNet to produce a sentiment score for each word and then used the scores along with other features to identify the stance in the SemEval stance dataset. Similarly, Krejzl and Steinberger [2016] used surface-level, sentiment and domain-specific features to predict the stance on that dataset. Those studies found that the use of sentiment in conjunction with other features helps in stance detection but not as the only dependent feature. A more recent study [Aldayel and Magdy, 2019] reveals that sentiment and stance are not highly aligned, and therefore the sentiment polarity cannot solely denote stance towards a given topic. They compared the distribution of sentiment and stance with respect to each topic, and found that sentiment fails to detect the real stance toward a topic, for instance, the negative sentiment does not match the "against" stance, and vice versa.

In this thesis, the sentiment polarity is used for detecting the attitude of news articles and the emotional state, which could partially reflect the viewpoints that are expressed in the news articles. Taking account of the disparate alignment between the stance and sentiment, the future work of this thesis could also investigate the stance in the news article separately, and merge this with sentiment polarity to enhance the detection of news attitudes.

2.3 Journalistic Framing on Climate Change

Although opinion mining is able to investigate the attitudes of newspaper articles reporting on climate change either '*positively*' or '*negatively*', the results only indicate the attitudes in a binary way. In a polarised media environment, partisan media publications intentionally frame news to advance, for example, certain political agendas [Jamieson et al., 2007, Levendusky, 2013]. Understanding news framing is helpful as it not only tells us whether a piece of news is positive or negative, but also investigates how the article is structured to promote a certain side of the political spectrum [Liu et al., 2019a].

News media has a democratic function to inform the public [Opperhuizen et al., 2018], and also interprets knowledge into the expression of popular discourse so that complex issues can be understandable to the public [Allan et al., 2000]. As Ungar [2000] stated, "*Science is an encoded form of knowledge that requires translation in order to be understood*". However, many studies have addressed the transmission failures from the scientists to the media [McComas and Shanahan, 1999, Ungar, 1992] and the media to the public [Stamm et al., 2000, Wilson, 2000]. Specifically, scientific findings normally come with follow-up experiments to either support or rebut the initial study. However, journalists cover initial studies far more often than follow-ups, and rarely inform the public when initial studies are disconfirmed [Dumas-Mallet et al., 2017]. For instance, an initial study claimed that a genetic factor was associated with depression when subjects were exposed to stressful life events [Caspi et al., 2003]. This finding was widely covered by 50 newspapers, and two subsequent studies confirmed the same association; however, newspapers never covered the 11 subsequent studies that failed to replicate this genetic association [Caspi et al., 2003]. Finally, a meta-analysis was published in 2009 [Risch et al., 2009] contradicting the initial finding, but was covered by only four newspaper articles. This is problematic when newspapers preferentially cover initial studies rather than subsequent observations, in particular those reporting null findings. In fact, it has been suggested that selecting intriguing story slants or news hooks would be the most crucial decision that journalists make for news reports, because such a selection gives meaning to events and issues, especially when they instigate reader's attention and interest [Zillmann et al., 2004]. This phenomena is also known as journalistic framing [Zillmann et al., 2004], in which journalists tend to give emphasis to certain aspects and downplay others, to capture and retain the readers' attention to the news [Gitlin, 2003, Tankard Jr, 2001].

Some research [Price et al., 1997, Gamson et al., 1992] has suggested why journalistic framing is so prevalent on influencing public opinions. On the one hand, even when faced with a diverse selection, news consumers are more likely to choose a news article which is most aligned with their pre-existing ideologies [Garrett, 2009]. For instance, a previous experiment [Knobloch-Westerwick and Meng, 2009], which recorded the time each participant spent looking at pro and con articles about social issues, found that participants spent 36 percent more time reading articles that agreed with their point of view. Moreover, this experiment also recorded the click rate to online articles, finding that participants reading just the abstract and topic had a 58 percent chance of choosing articles that supported their views, as opposed to a 42 percent chance of choosing an article that challenged their view. Thus, attracting attention from news consumers is salient for newspapers to survive in competitive markets, as the commercial pressure on media has increasingly dominated the institutional rules of news media [Štětka, 2013]. This is known as Market-Driven Journalism [Daniel, 1995], in which news readers are transformed into customers, news into products, and news circulation into markets. Meanwhile, such commercial pressures also influence how events and issues are packaged and presented by journalists, where news may be sensationalised to gain attention, expressed in binary terms to appeal to existing viewpoints, and where the selection of news stories and angles affect the way in which news consumers understand those events and issues. For instance, these selections could result in an 'echo chamber' [Jamieson and Cappella, 2008], where biased news is reinforced and repeated, until most people assume that some extreme variation of the story is true [Barberá et al., 2015], and have a trivialising and corrosive effect on public opinion and expression.

On the other hand, news media typically presents diverse accounts of news stories, and different publications present different angles on the same event. For instance, previous research [Jiang et al., 2017] found that articles studied in The Guardian generally had a positive attitude towards the Copenhagen Summit in 2009; however, most news reports from The Times expressed negative orientation to the same event. This is also because editorial decisions in newspaper articles are influenced by diverse forces and ideologies, such as political orientation. For instance, British newspapers, such as The Daily Mail, The Sun and The Daily Telegraph, were overwhelmingly in favour of Brexit, even though they are right-wing, with four times as many readers and anti-EU stories as their pro-remain rivals, whilst the Times and Guardian were not in favour [6]. Although newspapers have suffered from the rise of social media such as Facebook and Twitter, research[7] has found that the press still sets the agenda, where the newspapers lead on issues and broadcasters follow. Consequently, newspapers reflect the views of news consumers or influence votes [Reeves et al., 2016]. For example, the Sun claimed it was responsible for the unexpected Conservative general election victory of 1992.

To address this, some research [Wilkins, 1993, Jansen, 1981] has focused on the cultural and philosophical systems that affect news coverage. Bennett [1996] suggested that the content of news is affected by three normative orders that individual journalists must contend with: political norms (the idea that the proper role of the mass media is to provide the citizenry with political information that will lead to enhanced accountability on the part of elected officials); economic norms (the constraints on journalists working within a capitalist society in which reporting must be both efficient and profitable); and journalistic norms (objectivity, fairness, accuracy, balance). In response to the recent popularised term "*fake news*", which is a form of news consisting of deliberate misinformation, disinformation or hoaxes spread via traditional news media or online social media but more recently more likely to be a term used

[6] https://www.theguardian.com/media/2016/jun/24/mail-sun-uk-brexit-newspapers

[7] https://blog.lboro.ac.uk/crcc/eu-referendum/uk-news-coverage-2016-eu-referendum-report-5-6-may-22-june-2016/

by politicians to discredit mainstream journalism which criticises them, there are increasingly studies focusing on the identification of professional journalistic norms in news articles such as accuracy, and a sincere disposition towards truthfulness in newsgathering and reporting, especially in relation to journalistic framing of issues and events [Harrison, 2019, Zillmann et al., 2004].

Framing research has identified different types of news framing. Event-specific frames typically apply to unique events or issues, which differ from case to case [Reese et al., 2001, Davis, 1995, Jasperson et al., 1998]. For instance, a story about the Korean Airlines flight which was shot down by a Soviet interceptor in 1983 was framed as moral outrage, whereas the Iran Air flight which was shot down by a missile of U.S. Navy in 1988 was framed as a technical problem [Kuypers, 1997]. These event-specific frames, on the one hand, help to determine the importance of the events by the amount of news coverage. For instance, the New York Times printed 286 stories and the Washington Post printed 169 stories during the two-week period following the Korean flight shootdown; however, both publications printed only 102 and 82 stories respectively during the two-week period after the Iran flight shootdown. Although the two events had similar consequences, both newspapers were more likely to report stories about the Soviet interceptor in a background of the Cold War. On the other hand, the event-specific frames also have different narratives about these events. For example, the term "*attack*" was used 99 times and 66 times during the two-week period after the Korean flight shootdown in the New York Times and Washington Post respectively. However, the term "*attack*" was used only 30 times and 24 times in the above two publications during the two-week period after the Iran flight shootdown. Other frames are repeatedly and consistently employed, and permeate much of the news [Price et al., 1997, Semetko and Valkenburg, 2000]. It has been observed that U.S. news uses so-called conflict, economic-consequences, human-impact, and morality frames with considerable regularity. For instance, research has found that news reporting about the issue of unemployment focuses on vivid examples of people who have lost their jobs, whilst failing to link unemployment to any broader social, economic, or political processes [Price et al., 1997]. News readers thus would tend to make personal attributions (e.g., people are responsible for their poverty) rather than systemic attributions (e.g., poverty is because of institutional conditions).

These frames shape public attitudes on a variety of topics, and climate change is no different. For instance, conservative news outlets in the U.S., such as Fox News, are blamed for Republican climate denial by news commentators[8], as they have been found to disseminate misinformation on climate science, for instance, '*An analysis of cable news climate coverage finds Fox News 28% accurate, CNN 70%, and MSNBC 92%*'. [9]. Specifically, a report [Huertas and Kriegsman, 2014] was conducted to investigate the three most widely watched cable news networks in the U.S. (i.e., CNN, Fox News, and MSNBC) and their coverage of climate change, and found that 72% of 2013 climate science-related segments contained misleading statements in Fox News. CNN was the second, with about a third of segments containing misleading statements. MSNBC had the highest accuracy, with only 8% of segments containing misinformation. Furthermore, a 2012 UCS report[10] also found that the Fox's climate change coverage accuracy was actually improved from 7% to 28% in 2013, also most of the inaccurate segments on CNN were from debates featuring guests who reject aspects of established science because such networks try to be 'balanced' in their climate reporting. In the report, 38% of Americans watch cable news, and Fox News has the largest share of the audience compared with the others. This potentially explains why Americans are poorly informed about climate change. For instance, only two-thirds of Americans accept that climate change is occurring, and less than half of Americans recognise that it is largely due to human activities. This is a stark contrast to the 97% climate scientists consensus on human-caused global warming [Cook et al., 2016]. Such unbalanced narratives in climate change reports are also prevalent in the UK news media. For instance, the BBC has been accused of being the UK version of Fox News[11], and also research [Lewis and Cushion, 2009] has found that BBC relies heavily on sources from politics and business, but relatively less on academics and scientists.

Although journalistic framing brings increasing concerns, as it can enhance polarisation [Stroud, 2010] and perceptual biases [Barnidge et al., 2020], research analysing how to identify and evaluate these frames on a large scale is still lacking. Understanding and detecting journalistic framing in news media, and the influences on the diversity of perspectives of news media toward climate change, would potentially all lead to improved understanding of how

[8] https://www.theguardian.com/environment/climate-consensus-97-per-cent/2013/aug/08/global-warming-denial-fox-news

[9] https://www.theguardian.com/environment/climate-consensus-97-per-cent/2014/apr/08/fox-news-28-percent-accurate-climate-change

[10] https://www.ucsusa.org/resources/got-science-not-news-corporation

[11] https://www.theguardian.com/environment/climate-consensus-97-per-cent/2014/feb/27/bbc-false-balance-fox-news-global-warming

news media plays a role in climate change mitigation. Large-scale user studies are urgently needed in order to better understand how news consumers' beliefs on climate change are affected by the polarised and framed news articles. This thesis also explores how journalistic framing influences news coverage of climate change on a large scale. It focuses on the climate change-related entities, events and issues, investigating what are the attitudes towards them, whether these attitudes are polarised or balanced, and how these attitudes are framed and presented.

2.4 Tendentious Framing Detection

Journalistic framing is a subtle form of media manipulation which particularly gives emphasis to certain aspects and downplays others in news stories [Gitlin, 2003, Tankard Jr, 2001]. Detecting journalistic framing can offer insight into the selections and interpretations journalists make when framing a story, which can also define the nature of the debate and suggest to an audience how an event or an issue can be interpreted [D'Angelo and Kuypers, 2010]. This thesis aims to detect journalistic framing which is manipulated in the climate change related news. This section first looks at literature about how journalistic framing can be detected in news articles, and then discusses what types of framing could be extracted regarding in particular the climate change aspect.

2.4.1 Traditional Methods for Framing Detection

Approaches to detect journalistic framing have been analysed in many studies, and typically depend on pre-defined framing categories that are used in news stories. For instance, a previous study started with loosely theoretically defined frame categories to serve as guidance for the extraction of more specific frames through a grounded analysis, which aims to identify all the possible frames [Gamson et al., 1992]. Others focused on more specific frames directly, measuring the frequency of certain frames occurring in a given text. Semetko and Valkenburg [2000] investigated the news framing of European politics based on five pre-defined frames, namely conflict, human interest, morality, economic consequences, and responsibility. The frequency of frames was calculated by a series of questions that annotators had to answer. However, in the social sciences, framing is analysed traditionally by developing an extensive codebook of frames, reading large numbers of articles, and manually annotating them for the presence of frames in the codebook [Baumgartner et al., 2008, Terkildsen and Schnell, 1997]. Thus, detecting frame categories through qualitative analysis of a large sample of text would cause several difficulties (e.g., it is time-consuming and costly) [Reese, 2007], and computational linguistic methods are needed to automatically detect formalised journalistic framing on a greater scale [Card et al., 2015].

Several studies have built automatic methods to detect journalistic frames by combining probabilistic topic models. For instance, Tsur et al. [2015] investigated the journalistic frames based on a four-year set of public statements issued by members of the U.S. Congress, by combining probabilistic topic models with time series regression analysis. Nguyen et al. [2015] introduced a Hierarchical Ideal Point Topic Model to focus on policy issues, framing and voting behaviour through the relationship between Tea Party Republicans and "establishment" Republicans during the 112th U.S. Congress. Those works focused on event-specific frames and revealed particular opinions on certain events or issues. However, the abstraction of generalized frames which allows comparison across many social issues is missing.

To address this, recent research [Card et al., 2015] established a large-scale dataset of frame annotations, namely the Media Frame Corpus (MFC). It includes three topics: tobacco, immigration and same-sex marriage, which were classified into 15 generic media frames defined by the Policy Frames Codebook [Boydstun et al., 2014]. Several framing detection studies have been implemented based on the MFC. For instance, Naderi and Hirst [2017] applied deep learning models to represent the meaning of frames and classify those 15 frames at the sentence level in news articles. More recently, Liu et al. [2019a] detected frames in news headlines which related to U.S. gun violence by implementing a language model. Although the news articles from MFC involved several controversial social issues, they do not include climate change aspects. Meanwhile, the 15 framing categories of the MFC dataset were annotated based on the *Policy Frames Codebook* that mainly focused on political aspects. However, this thesis particularly looks at the balance of news attitudes, the bias of reporting styles, and eventually how these affect news consumers' climate change beliefs. Thus, news frames which are specific to the fairness and tendentiousness of news articles, specifically need to be identified in this study. Furthermore, automatic detection of journalistic framing with respect to climate change news articles has never been studied before as far as we know.

2.4.2 Tendentious Framing and Hyperpartisan Framing

This research looks at two framing styles, namely **Tendentious** [Harrison, 2019] and **Hyperpartisan** [Potthast et al., 2017]. Tendentious framing can be defined as news reports persuading their audience in a campaigning and universalistic way, and events are typically explained as a direct advocate of a specific cause. In the tendentious news reports, stories and their evidence are manipulated to support a particular viewpoint, and others that do not agree are downplayed or ignored. Although many viewpoints could be expressed in the tendentious style, viewpoints are explicitly advocated towards one side of an event. Essentially, these news articles are taking advantage of emotional elements to exaggerate or understate human interest stories, and might even seem unpleasant to the audience [Harrison, 2008]. Detecting tendentiousness in news articles is essential to understand why the presented news stories reflect particular opinions and attitudes, which can heavily influence the perspectives of readers. Furthermore, increased partisanship in the news media, which can result in misunderstanding and misuse of facts, is a factor in changing individuals' voting preferences [Gentzkow, 2016], and has even led to ethnic violence [Minar and Naher, 2018]. This type of news, which expresses an extremely one-sided opinion or unreasoning allegiance to one party, has been recently defined as hyperpartisan news [Potthast et al., 2017]. The characteristics of hyperpartisan news are highly similar to the tendentious framing in the newspaper articles. For instance, they both use emotional and inflammatory language to frame the context of a news story. Furthermore, they are typically extremely one-sided and often riddled with untruth [Potthast et al., 2017]. However, hyperpartisan news seems to be mainly focused on politics, while tendentiousness is found equally in all types of news and is not necessarily so extreme.

Table 2.1 depicts two news snippets in the manner of hyperpartisan and tendentious framing style respectively. Specifically, both snippets use emotional phrases (e.g. *braggadocios*, *unrealistic* and *panic*, etc) to convey a one-sided point of view, and also both employ sarcasm to frame the news events (e.g. "*the best, the hugest, the most competent...*" and "*It's OK everybody. Panic over...*"). This provides an added challenge, as detecting such subtle discourse from newspapers is considered too tricky to handle even for people sometimes [Maynard and Greenwood, 2014]. However, the differences between hyperpartisan and tendentious news are in the detail. The hyperpartisan news, as its name implies, can be found mostly in political news articles (e.g. elections, campaigns, etc). The tendentious style, on the other hand, advocates a

Framing style	
Hyperpartisan	*Donald Trump ran on many braggadocios and largely unrealistic campaign promises . One of those promises was to be the best , the hugest , the most competent infrastructure president the United States has ever seen . Trump was going to fix every infrastructure problem in the country and Make America Great Again in the process . That is , unless you 're a brown American . In that case , you 're on your own , even after a massive natural disaster like Hurricane Maria.* (12th Oct 2017, The Bipartisan Report)
Tendentious	*It's OK everybody. Panic over. I'll bet I'm not the only one to be feeling all tickety-boo at the news that human-caused climate change probably isn't anything to get worked up about and that we can all go back to worrying about deforestation, rampant over-consumption of the world's natural resources, the collapsing ocean ecosystem and whether or not Donald Trump's hair might be real. How fortuitous, too, that the news comes as more than 190 governments from around the world are preparing to head to Paris next month to agree a new global deal to cut greenhouse gas emissions. In case you hadn't read the big news.* (16th Oct 2015, The Guardian)

Table 2.1: Examples of news snippets from two different reporting styles.

specific cause to explain events in terms of human interest stories, which could be found in most kinds of news articles. In this context, this thesis assumes that the hyperpartisan news is a subset of the tendentious one, and is typically more extreme, and specific to politics. Apart from that, hyperpartisan news has a high similarity to tendentious news in terms of discourse.

2.4.3 Transfer Learning for Tendentious Framing Detection

Deep Learning-based approaches have achieved state-of-the-art performances on many downstream Natural Language Processing (NLP) tasks. Although these work very well on large data sets, on small data sets they fail to achieve significant gains. Meanwhile, annotating a news corpus manually is both time and labour-intensive. In order to address such issues, the concept of transfer learning can be exploited to improve the performance of the deep learning model, by training one model from one dataset and fine-tuning it based on a similar domain dataset [Howard and Ruder, 2018].

Transfer learning is applicable when there is a lack of labelled data or when the data becomes quickly outdated, but when knowledge could still be obtained from similar tasks or domains [Weiss et al., 2016]. Calais Guerra et al. [2011] implemented transfer learning that measured the bias of social media users towards a topic, then transferred the user bias to textual features to analyse sentiments in real-time. Li et al. [2010] built a model that learnt the linguistic expressions and sentiment terms in a movie reviews dataset, and transferred this knowledge to target domains such as software reviews and political blogs. Blitzer et al. [2007] investigated domain adaptation for sentiment analysis by implementing a structural correspondence learning algorithm to reduce relative error due to adaptation between domains.

Recently, a hyperpartisan dataset has been openly made available in the International Workshop on Semantic Evaluation 2019 (SemEval 2019) task 4: Hyperpartisan News Detection.[12] The release of this corpus makes it possible to build machine learning models to automatically detect partisanship in news articles [Jiang et al., 2019a]. However, this is not specific to climate change, but covers all kinds of political news, and to the best of our knowledge, no tendentious newspaper corpus exists so far. This means that building a tool to detect tendentiousness in climate change articles is tricky.

This thesis uses a crowdsourcing platform to build a tendentious news corpus about climate change (see Chapter 7). The idea is to get away with quite a small corpus here, saving time and effort, by taking advantage of transfer learning on the large-scale hyperpartisan news corpus, when it comes to building our deep learning models.

2.4.4 Deep Learning for Tendentious Framing Detection

Deep learning methods have already made impressive advances in fields such as computer vision [Krizhevsky et al., 2012], pattern recognition [Wang et al., 2019] and also NLP [Yang et al., 2016b]. Traditionally, machine learning approaches based on shallow models (e,g. SVM and Logistic Regression) are trained on high dimensional, sparse and hand-crafted features (e.g., TFIDF and BoW vectors). Learning in such a high-dimensional but sparse dataset could be limited in terms of computation and memory, and thus typically needs dimensional reductions. For instance, Huang and Yates [2009] implemented Latent Semantic Analysis (LSA) to reduce the sample complexity for

[12] *https://pan.webis.de/semeval19/semeval19-web/*

the sequence labelling task, while Väyrynen et al. [2007] used clusters induced from distributional similarity from vector space to reduce high dimensionality.

On the other hand, hand-crafted features require human assessments and are sometimes time-consuming and costly [Young et al., 2018]. For instance, named entity recognition (NER) tasks typically require both lexical and syntactic knowledge, such as part of speech (PoS) and chunk tags, prefixes and suffixes, or external gazetteers, all of which need to be extracted and selected carefully [Wu et al., 2018]. Deep Learning methods, however, enable multi-level automatic feature representation learning without human assessing, and neural network-based dense vector representations have been producing state-of-the-art performance on various NLP tasks, such as machine translation [Koehn, 2020] and text classifications [Yang et al., 2016b].

In communication research, manual identification of journalistic framing is challenging, due to the large volume of online news data along with the growth of news media digitisation. Furthermore, the detection of journalistic framing has a high level of complexity and often requires careful investigation of nuances in news coverage, which is time-consuming [Liu et al., 2019a]. In the field of NLP, existing opinion mining techniques fall short of addressing the nuances needed for framing detection, which requires the detection of perspectives beyond just positive and negative, such as the tendentiousness [Harrison, 2019] and the partisanship [Potthast et al., 2017] of news reporting.

For automatically detecting framing, distributional vectors or **Word Embeddings** are first introduced. Specifically, word embeddings try to capture the characteristics of the neighbours of a word by following the distributional hypothesis that “words with similar meanings tend to occur in similar context” [Mikolov et al., 2013]. Word embeddings are often used as the input in a deep learning model, and are typically obtained by optimising an auxiliary objective in a large unlabelled corpus, such as predicting a word based on its context, or vice versa [Mikolov et al., 2013]. Such characteristics allow word embeddings to have the ability to capture general syntactic and semantic information, and therefore can measure text similarity and analogies effectively based on their dense dimensionality [Young et al., 2018].

Morphological information, such as the composition of letters, can also be useful for tasks such as POS-tagging and named entity recognition (NER), but such features are typically not used in traditional word embedding models. Also, the creation of word embeddings requires a dictionary to be built for the entire vocabulary, but this is problematic when there are unknown words or out-of-vocabulary (OOV) issues. Thus, several studies [Kim et al., 2016, Dos Santos and Gatti, 2014, Santos and Guimaraes, 2015] have focused on character-level embedding, and better results on morphologically rich models are reported. The OOV issue is also addressed by character level embeddings, as each word is considered as a composition of individual letters.

Apart from character level embedding, traditional word embeddings have developed and expanded in recent years, especially with the emergence of contextualised word embeddings in the era of NLP pre-trained language models [Qiu et al., 2020]. Traditional word embeddings generate a context-free and fixed representation for each word in the vocabulary. For instance, the word "apple" has the same vector representation for the meanings "eat apple" and "the Apple company"; however, the context completely changes the meaning of "apple" in a sentence. For this reason, recent pre-trained language models (e.g. ELMo [Peters et al., 2018] and BERT [Devlin et al., 2018b]) utilise bidirectional approaches to guard against context-free issues, and such models have achieved state-of-the-art performance in many NLP tasks [Jiang et al., 2019b, Alsentzer et al., 2019, Wiedemann et al., 2019, Peters et al., 2018, Devlin et al., 2018b]. Table 2.2 lists some popular word embeddings.

Word Embeddings	**Reference**
Word2Vec	[Mikolov et al., 2013]
GloVe	[Pennington et al., 2014]
Fasttext	[Bojanowski et al., 2017]
Flair	[Akbik et al., 2019]
ELMo	[Peters et al., 2018]
BERT	[Devlin et al., 2018b]

Table 2.2: List of Word Embeddings

Following the popularisation of word embeddings and their ability to represent words as feature vectors in a distributed space, these abstract features can be used by different deep learning structures such as **Convolutional Neural Networks (CNN)** and **Recurrent Neural Networks (RNN)**. The effectiveness of CNNs has been demonstrated in many computer vision tasks [Krizhevsky et al., 2012, Sharif Razavian et al., 2014, Jia et al., 2014], while Collobert et al. [2011], Kalchbrenner et al. [2014], Kim [2014] also pioneered CNNs in the field of NLP.

In the text classification task, as shown in Figure 2.2, CNN models start with a look-up layer that aims to transform each word into a word embedding matrix with pre-defined dimensions *d*. In the convolutional layer, there are a

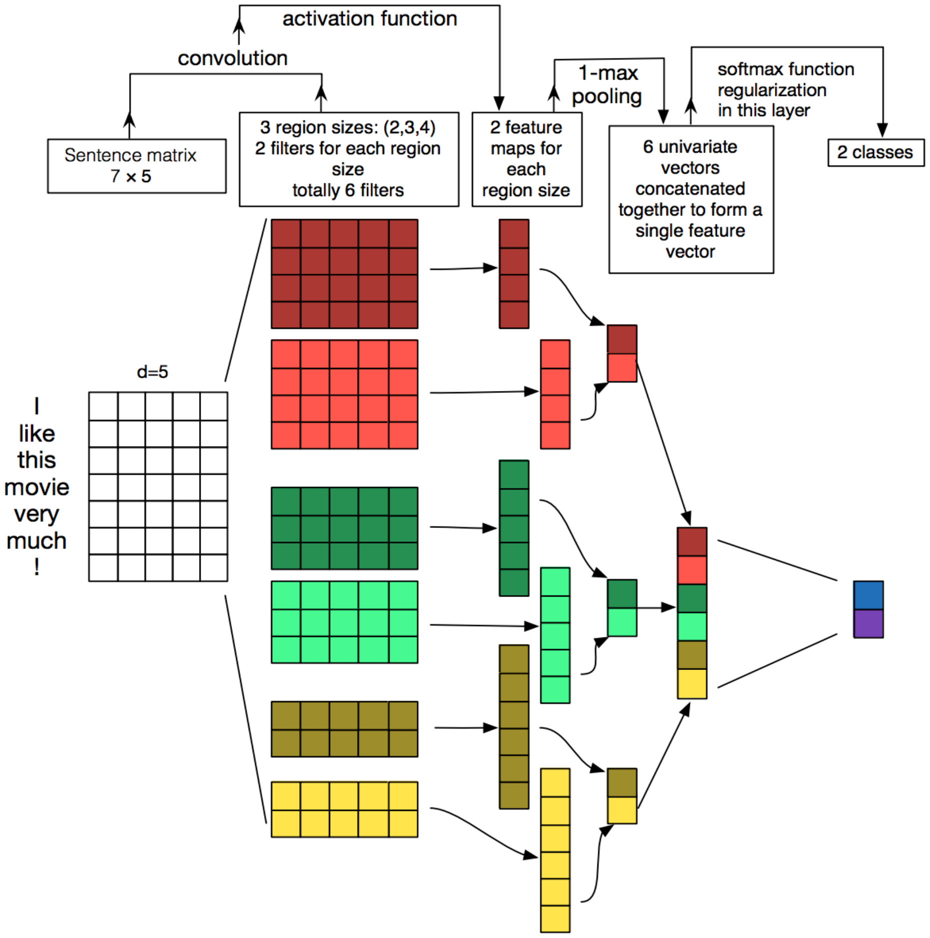

activation function
convolution
Sentence matrix
7 × 5
3 region sizes: (2,3,4)
2 filters for each region size
totally 6 filters
2 feature maps for each region size
1-max pooling
6 univariate vectors concatenated together to form a single feature vector
softmax function regularization in this layer
2 classes
d=5
I
like
this
movie
very
much
!

Figure 2.2: Convolutional Neural Network Architecture for Sentence Classification [Zhang and Wallace, 2015].

number of filters, also called kernels, of different region (or window) sizes which slide over the entire matrix. Each filter extracts a specific pattern of n-grams (i.e., 2,3,4 region sizes) and obtains feature maps accordingly. Then, a max-pooling layer typically is used to subsample the feature maps by applying max operation on each feature map. Regardless of the size of the filters, max pooling always maps the input to a fixed dimension of outputs, and also reduces the dimensionality from the feature maps while keeping the most salient n-gram features across the whole sentence. Finally, the outputs from the max-pooling layer are normally concatenated as input for a fully connected layer with a Softmax activation function, to make the final prediction.

The RNN is designed for processing sequential information, and has the ability to capture the inherent sequential nature present in language [Elman, 1990]. Words in a language develop their semantic meaning based on the context, such as the previous words in the sentence (e.g., the meaning of ‘dog’ and ‘hot dog’ are different). RNNs can model such contextual dependencies in

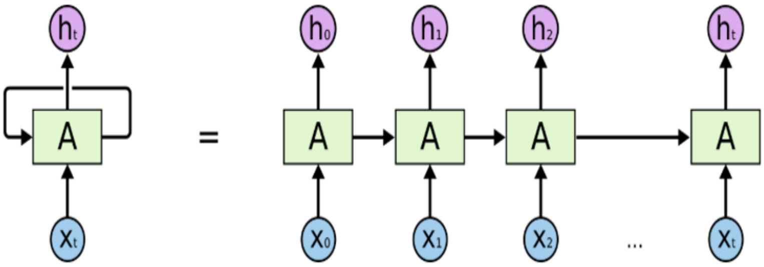


Figure 2.3: Recurrent Neural Network Architecture[13]

language and have provided stronger motivation for researchers to use RNNs over CNNs [Young et al., 2018]. Furthermore, they have the ability to model a variable length of text, including very long sentences, paragraphs and even documents. The basic RNN model is demonstrated in Figure 2.3, where x_t is taken as the input (normally these are embeddings of each word) to the network at time step t and h_t denotes the hidden state at the same time step. Thus, each h_t is calculated based on the current input and previous time step's hidden state in a fully connected layer. It can then use the final hidden state, a fixed length vector, to make the final prediction with Softmax activation.

However, such simple RNNs suffer from the infamous vanishing gradient [Hochreiter, 1991, Bengio et al., 1994] problem, especially when the context dependency becomes too large. For example, consider trying to predict the last word in the sentence "I grew up in China ... I speak fluent *Chinese*." The recent information 'speak' suggests that the last word should be a language, but if we want to know which language, we need the context of China from further back. This is the so-called long-term dependencies problem, and the RNN becomes unable to learn the connections between the information whilst the gap is growing. To address this, Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] and Gated Recurrent Units (GRU) [Cho et al., 2014] were designed by applying gated units which allow the error to back-propagate through an unlimited number of time steps [Young et al., 2018], and therefore handle the long-dependency issue. The applications of LSTM and GRU have demonstrated their superiority over the traditional RNN. However, in this work there were no concrete conclusions about which of the two gating units was better [Young et al., 2018].

2.5 Summary

To conclude, issues related to climate change events are popular in newspaper reports, because they make for dramatic and enticing stories about which people tend to have strong views. This thesis first looks at the attitudes/opinions of different newspaper articles toward climate change issues, and investigates how these attitudes are different in the reporting of different newspapers about certain aspects of climate change, by using opinion mining techniques.

Furthermore, as climate change reflects the general public's concern and social opinion, news reports are typically framed in a dramatic manner rather than just being presented as factual events [Beattie and Milojevich, 2017]. The controversiality and popularity of climate change issues in the news, which lead to strong polarisation of opinions, make it an ideal topic for the study of tendentiousness. To address this, this thesis investigates different deep learning architectures for detecting journalistic framing, especially in tendentious framing and hyperpartisan framing. Since the annotation of tendentious news is costly, the work starts with establishing deep learning models that could accurately distinguish the hyperpartisan framing from existing large hyperpartisan datasets. Based on the similarity of two journalistic frames, it then uses a transfer learning approach to identify tendentiousness with a relatively smaller tendentious dataset acquired by crowdsourcing. In order to evaluate the performance of learning models, it explores different combinations of word embeddings (e.g., BERT and ELMo) and deep learning architectures (e.g., CNN, RNN and their variants) in the task of journalistic framing detection.

Chapter 3

Opinion Mining on Climate Change News

As discussed in the previous chapter, the news media may not produce unbiased news stories, and different publications present different angles on the same event. In order to automatically identify the attitudes in the climate change news (i.e., RQ1), this chapter investigates how different publications differ in their approach to stories about climate change, by examining the opinions and topics presented. To understand these attitudes, this chapter finds opinion targets by combining Latent Dirichlet Allocation (LDA) with SentiWordNet, a general sentiment lexicon. This chapter also describes the creation of corpus of news articles and the use of LDA to extract topics from it, which contain keywords representing the opinion targets. Then, sentiment is assigned using SentiWordNet, before regrouping the articles based on topic similarity. The work described in this chapter aims to automatically identify different attitudes of news publications toward certain climate change issues by combining LDA and opinion mining.

3.1 Problem Definition

Editorial decisions in newspaper articles are influenced by diverse forces and ideologies. News publications do not always present unbiased accounts, but typically present frames reflecting opinions and attitudes which can influence the readers' perspectives [Spence and Pidgeon, 2010]. Climate change is a controversial issue in which this kind of framing is very apparent.

Opinion mining is typically implemented on short documents such as Twitter [Pak and Paroubek, 2010, Agarwal et al., 2011] and customer reviews [Pang et al., 2008, Shelke et al., 2017]. Since such documents contain a short piece of information, the majority of them are therefore considered as conveying a single topic due to this length limitation [Giachanou and Crestani, 2016b,a], and the complexity of their context thereby could be simplified as only one opinion per tweet. However, newspaper articles have diverse context length (i.e., the huge variation in word counts in news articles), so their content is much more complicated than other types of sources, especially because the topics of the news are normally varied (i.e., talking about different aspects of an event). For instance, a variety of smaller topics might be discussed in the context of a particular climate change issue in a news article. Thus, it is necessary to understand what the target of the opinion is in each case, i.e. which aspect of climate change the opinion is about.

This chapter examines a set of articles about climate change in four UK broadsheets between 2007 and 2016. Quality newspapers, which typically offer intellectual analysis and balanced viewpoints, play a crucial role as agenda-setters and as mirrors of public concern in relation to climate change [Barkemeyer et al., 2017]. Thus, this chapter aims to investigate how quality newspapers differ in their reporting about the climate change issue, by examining their topics and opinions. It is impractical to identify topics and analyse all the opinions about them manually in this large set. In this chapter a topic modelling method is therefore developed to generate topics using LDA, and the articles are then clustered into groups with similar topics. Then it performs opinion mining on each cluster by using a sentiment lexicon, SentiWordNet [Baccianella et al., 2010], in order to investigate the opinions, and how they differ in the various news publications. Consequently, the method described in this chapter, which combines LDA with SentiWordNet, provides an overview of how news attitudes differ in various climate change topics in the news media.

3.2 Methods

This section first presents the data collection and corpus statistics. Second, the pre-processing methods are introduced. Third, the implementation of the LDA model is presented in detail. Finally, SentiWordNet is introduced for assigning sentiment labels to each news article.

3.2.1 Data Collection

Broadsheets or quality newspapers have long been argued to stand out in providing a larger amount of coverage on political issues and also to have a higher agenda-setting impact than tabloid newspapers [Carvalho and Burgess, 2005, McCombs, 2018]. This thesis also assumes that different broadsheets tend to have different viewpoints since their political orientations are different. Thus, the initial corpus consists of 11,720 newspaper articles which were collected from four UK broadsheets with different political orientations: *The Guardian* (a left-leaning quality newspaper); *The Independent* (a centre-left quality newspaper); *The Telegraph* (a right-leaning quality newspaper); and *The Times* (a centre-right quality newspaper). The newspaper articles were selected from the digital archive LexisNexis[1] with the search query "Climate" OR "Climate Change" between 2007 and 2016. Some of the news articles occur more than once (i.e., same titles and contents) – the articles which are duplicated are removed. The final corpus contains 7,429 newspaper articles, with statistics as shown in Table 3.1. Although there are issues of imbalance in this dataset (e.g., The Times has only around a fifth of the number of articles that The Guardian has), this has a minimal negative effect on the method used in this thesis for two reasons. 1) Since the opinion mining is purely based on SentiWordNet lexicon, all the sentiment labels are assigned based on the term frequencies, thus no classifiers are trained. Also, the sentiment scores are normalised to deal with the imbalanced number of articles. 2) The LDA topic model is fully unsupervised and takes account of the entire corpus as a whole, thus the topic distributions are equally assigned to each article regardless of its source.

Publishers	**num. of articles**
Guardian	2874
Independent	2651
Telegraph	1349
Times	555

Table 3.1: Statistics per publisher.

[1] https://www.lexisnexis.com/uk/legal

3.2.2 Pre-Processing

In the pre-processing, all the news articles are tokenised and stemmed using the NLTK[2] toolkit, and stop-words, which typically refer to the most common words (e.g., 'and', 'is' and 'are', etc), are removed as they do not indicate useful semantic content and are therefore not considered as opinion targets when generating topics.

The NLTK Part Of Speech (POS) tagger is used to annotate all the words. To ensure the generated topics are related to certain events/issues, which are often represented by proper nouns (e.g., Copenhagen Summit), the quality of the generated topics is manually checked and other POS (e.g., verbs, prepositions, adjectives, etc.) are removed, keeping only the nouns for the LDA model to generate sentiment targets. However, every word matters for opinion mining, thus all words are retained except for the stop words which are already removed from the news articles after the topics are generated.

3.2.3 LDA Model

Climate change is a broad topic which typically contains many subtopics (e.g., pollution, carbon emission, etc). The contents of news articles therefore typically refer to different kinds of subtopics when talking about climate change. In order to identify various aspects of the climate change issue mentioned in the newspaper articles, and the attitudes toward such issues, this research implements LDA for automatic topic extraction on a large volume of news articles.

LDA is a probabilistic model for grouping topics in documents according to a predefined number of topics. However, such a predefined number will result in limited word correlation with topics if the number of topics is predefined incorrectly. A too large or too small number of predefined topics might cause the inaccurate grouping of topics during training of the LDA model [Arun et al., 2010].

Typically, the optimal number of topics is found by maximising the distance between the different topics, thus ensuring that the topics are as distinct as possible from each other [Arun et al., 2010]. Different methods are used to determine the number of topics in LDA such as the perplexity on a held-out set [Song et al., 2020] or topic coherence [Jiang et al., 2020]. In order to find out how closely related the documents in a cluster are, and also how distinct

[2] https://www.nltk.org/

(or well-separated) a cluster is from other clusters, therefore this study treats the topics as clusters, and applies the Silhouette Coefficient to determine the number of topics for the LDA model. Silhouette is a well-established measure for cluster validation that considers both how similar each object is to its own cluster (cohesion) and how different it is to other clusters (separation). This method has been previously used for finding the optimal number of topics [Panichella et al., 2013, Ma et al., 2016]. The Silhouette Coefficient equation (Eq 3.1) can be described as follows:

$$Sil = \frac{b - a}{max(a, b)} \tag{3.1}$$

where a is the mean distance between a point and other points in the same cluster, and b is the mean distance between a point and other points in the next nearest cluster. In the Silhouette analysis [Ma et al., 2016], Silhouette Coefficients (Sil) close to +1 indicate that the samples in the cluster are far away from the neighbouring clusters. In contrast, a negative Silhouette Coefficient means that the samples might have been assigned to the wrong cluster.

In this study, the analysis was run repeatedly on the entire data set with a different number of topics (0-30) and the silhouette value for each number of topics was added to the plot in Figure 3.1. The highest Silhouette Coefficient score orrcurs when the number of topics reaches 20. After that, the Sil score starts to decrease whilst the topic number grows. Thus, this study uses 20 topics as the hyperparameter to the LDA model finally.

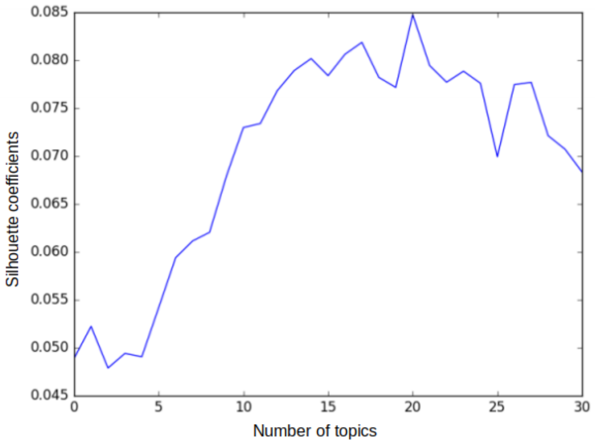


Figure 3.1: Silhouette analysis for LDA model

After the number of topics has been determined, the LDA assigns keywords to one of the topics of the news article, based on the probability of the keywords occurring in the topics. This assignment also gives topic representations of all the articles. This assignment was repeatedly updated for 50 iterations to generate both topic distribution in the articles and word distribution in the topics. For each topic in the LDA model, the top 10 keywords are selected (i.e., the words which have the 10 highest probabilities for that topic) with their distribution to represent the corresponding topic (see Table 3.2).

Topic_ID	**Keywords (Weights)**
Topic 1	land (0.084), world (0.079), food (0.031),...
Topic 2	science (0.098), year (0.053) , time (0.03)...
Topic 3	world (0.029), car (0.021), weather (0.018)...

Table 3.2: Example of topic list in The Guardian 2007. Each topic has 10 keywords with its weight which reflect how important a keyword is to that topic.

Each article is assigned to a set of topics, and each topic generates a set of keywords based on the vocabulary of the articles. After acquiring the topics from the LDA model, the bag-of-words model is converted into a topic-document combination, where each document then can be seen as a low dimension matrix (Table 3.3). It then selects the topic with the highest probabilities among the 20 topics (i.e., the topic which has the highest probability of occurring in the article compared with other topics) from each news article in the different news sources.

Articles	**Topic_ID**	**Probability**
Article 1	1	0.519842
Article 2	12	0.348175
Article 3	7,12	0.412394, 0.1492813
Article 4	2	0.249132

Table 3.3: Example of topic-document combination. For each article, it shows the topic ID numbers with their corresponding probabilities.

.

3.2.4 Applying SentiWordNet

To automatically annotate the articles with sentiment labels, SentiWordNet[3] was used, which contains roughly 155,000 words associated with positive/negative (PN) polarity score and subjective/objective polarity score. For instance, the

[3] http://sentiwordnet.isti.cnr.it/

word ‘faithful’, as an adjective, has 0.625 positive score and 0.375 objective score in SentiWordNet (see Figure 3.2).

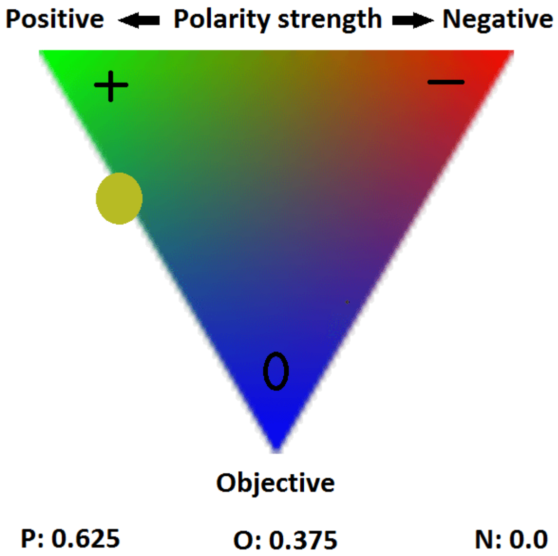


Figure 3.2: Example of SentiWordNet 3.0 online graphical representation of first sense of the word ‘faithful’ as an adjective [AL-Sharuee et al., 2017].

Sentiment labels are assigned based on the words associated with sentiment polarity scores in SentiWordNet from each article. Then, it is necessary to identify which articles have similar topics, especially across all publications. Here, it is assumed that those news articles are similar as their topics have similar or the same keywords, as well as similar opinion targets. Such assumption is based on the probability of the topics in the articles and also the probability of the keywords in the topics, since it selects the topic and keywords which all have the highest probabilities in the articles and topics.

Once there is a score for each article, the different attitudes of each news source on the same climate change issue can then be analysed. For this, the keywords in the topic lists in each news source in each year are manually checked, and those topics containing at least two of the same keywords, are grouped together. Specifically, every keyword in each topic ID from 2007 to 2016 in each news source is analysed, and the keywords which occur in each topic are extracted. Then, the topic IDs are extracted based on those keywords, and the IDs are grouped based on the topics that contain at least two identical keywords.

Although the partly manual aspect of this method has limitations of scalablility on a large dataset, the aim here is to show a case study about how the approach might be used to analyse the different attitudes expressed in the news about the same topic (i.e., see Section 3.3).The limitations will be discussed further in Section 3.4.

3.3 Results and Discussion

The four news publications were compared by analysing the clusters identified by the method. Although there are a few similar topics mentioned across all four news publications for some years, one example that stands out is the reporting by all four broadsheets of the Copenhagen Summit in 2009 (see Table 3.4). The clusters all contain the keywords "copenhagen" and "agreement", which refer to the Copenhagen Summit explicitly. This feature of shared keywords allows us to identify the main topics, which also can be seen as the sentiment targets, and was utilised to compare the different attitudes toward the same issue (Copenhagen Summit) between the four news sources. However, the keywords are mostly different between the sources in other years. For instance, some topics in The Guardian and The Times have large numbers of keywords such as "gas" and "energy" in 2012, but topics in The Telegraph in that year are associated with the keyword "wind", while The Independent has keywords like "government" and "investment". This means that there are no good shared topics in this case for which the method can compare attitudes.

Sources	**Topics**
Guardian	copenhagen, world, deal, agreement, summit, president, obama, china , action, treaty
Times	copenhagen, world, cent, deal, president, summit, agreement, conference, china, year
Telegraph	world, carbon, copenhagen, summit, deal, cent, agreement, energy, time, president
Independent	world, carbon, copenhagen, deal, cent, agreement, year, conference, cancun, government

Table 3.4: Topics in the year of 2009.

Figure 3.3 shows how sentiment differs between the reports about the Copenhagen Summit in 2009 in the four broadsheets. Table 3.5 gives also some examples of positive and negative sentences found. A manual check of a random selection of the relevant articles confirms the general tendency. Most of the articles used some negative words, such as "failures", "collapse", "drastic". However, Figure 3.3 indicates that the overall sentiment is relatively impartial to positive (the average sentiment score across all sources is +0.15).

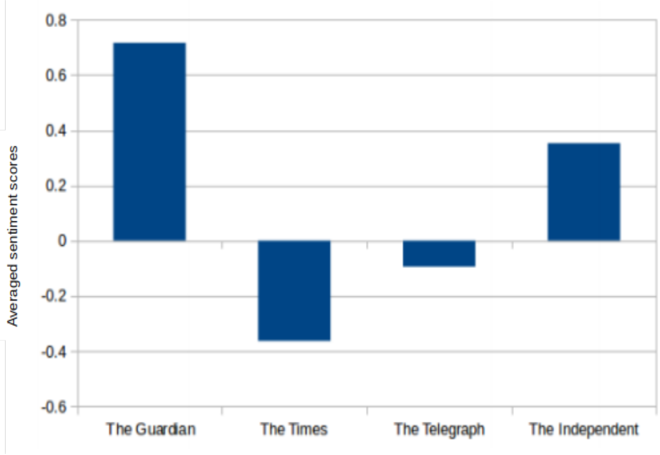


Figure 3.3: Attitudes of four news sources to the Copenhagen Summit in 2009.

The Guardian is the most positive, while The Times is the most negative. This study suspects that some of the keywords may be a bit misleading (e.g the word “agreement“ is typically positive, but here it is often being used in a neutral way), which might influence the sentiment analysis. Also, the overall sentiment orientation tends to be positive as the lexicon-based approach assigns sentiment scores without taking account of any syntactic rules (e.g., negation) for news articles.

Detected Sentences
Positive
China itself defended its crucial role in saving the Copenhagen conference from failure. (The Guardian, 28 Dec, 2009)
Don’t panic. Copenhagen really wasn’t such a disaster. (The Independent,15 Dec, 2009)
Negative
The move emerged from the chaotic Copenhagen conference on climate change. (The Telegraph, 21 Dec, 2009)
Copenhagen puts nuclear options at risk. (The Times, 23 Dec, 2009)

Table 3.5: Example sentences with sentiment polarity detected in the four news sources in 2009.

However, there are some clear indications that match the automatic analysis results. While The Guardian does have some quite negative reports about the summit, mentioning things like “catastrophic warming”, it also tries to focus on the hope aspect (“The talks live. There is climate hope. A bit. Just.”). The Independent tends also towards the positive, talking about leaders achieving ”greater and warmer agreement”. The Telegraph, on the other hand, plays more on the fear and alarmist aspect, talking about ”drastic action” and ”imminent dangerous climate change”, although also about positive steps towards the future. The Times, on the other hand, emphasises the role of honesty; although its overall tone is not overwhelmingly negative, it does mention repeatedly the fear and alarmist aspect of climate change and some of the negative points about the summit (for example that Obama will not be there).

3.4 Limitation and Future Works

Although the method could identify the similar topics across four news publications, the number of such topics is small. This is probably because different newspapers attached different levels of importance to most topics. For instance, even with similar topics (see Table 3.4), The Guardian focuses on the political 'deal' between U.S. and China, but The Telegraph, however, focuses on carbon emissions. A potential solution is that, instead of just using general keywords (e.g., 'climate' or 'climate change'), the search query could specify a particular topic/event (e.g., 'Copenhagen Summit'), or use the combination of both (e.g., 'climate change' and 'Copenhagen Summit').

The sentiment assignment is purely reliant on a simple lexicon-based method without considering syntactic rules (e.g., negation). Meanwhile, words with multiple senses are assigned with a fixed sentiment score by averaging scores from different senses [Hamouda and Rohaim, 2011]. For instance, the word 'agreement' is positive after averaging over all senses, however, in our case it refers to the event which should be more neutral. This could be solved by implementing word embedding techniques, which generate high quality word representation to preserve semantic relationships between words and their context, such as Word2Vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014], and ELMo [Peters et al., 2018]. Future work could implement deep learning methods to generate contextually sensitive word representation to capture the semantic information in the news article.

As discussed in Chapter 2, the sentiment polarity might only partially reflect the viewpoints that expressed in the news article, and there is a disparate alignment between the stance and sentiment. Thus the stance could also be merged with sentiment features for detecting the news attitudes in the future work.

3.5 Summary

This chapter has presented a method for combining LDA with opinion mining to detect the attitudes of four news publications towards climate change related issues. Since the topics of the news are normally varied (i.e., talking about different aspects of climate change), traditional methods are typically carried out manually, and therefore limited to small case studies. The aim, however, is to apply such techniques on a large scale and also minimize manual assessments.

The LDA model provides high interpretability for a large corpus by generating a set of topics automatically. Therefore, the chosen method here combines an LDA model with SentiWordNet to automatically extract topics from the articles, and then regroups these articles based on their topic similarity, followed by assigning a sentiment score for these groups of articles. Specifically, similar topics were identified between different news publications by utilising the probabilities of topics for each article, and also the distribution of keywords for each topic. The method assumes that the similar topics also indicate similarity of the opinion targets between articles based on their highest probabilities in topics and keywords. The experimental results demonstrate that the method is able to extract similar topics from different publications and to explicitly compare the attitudes expressed by different publications while reporting similar topics.

The limitations of this approach is that the number of shared topics between different news publications is small. Therefore, the comparison of attitudes is restricted to a limited number of climate change subjects. Also, the sentiment analysis is only based on a lexicon-based approach that did not consider syntactic rules. In the future work, this study will try to narrow down the search keywords to a specific event or issue in order to make the shared topics have a similar domain.

Chapter 4

ELMo Sentence Representation Convolutional Network

Although opinion mining is able to investigate the attitudes of newspaper articles reporting on climate change, it only tells us whether a piece of news is positive or negative. In a polarised media environment, partisan media publications may intentionally frame news to advance, for example, certain political agendas [Jamieson et al., 2007, Levendusky, 2013]. Therefore, understanding news framing is helpful as it also explains how the article is structured to promote a certain side of the political spectrum [Liu et al., 2019a].

The recently released hyperpartisan dataset[1] makes it possible to build deep learning models to automatically detect partisanship in news articles on a large scale. Since there is similarity between hyperpartisan framing and tendentious framing, this thesis takes the advantage of the similarity of hyperpartisan news and tendentious news, and implements a transfer learning approach to detect the tendentious news using the knowledge acquired from the hyperpartisan news. However, this means that the initial learning model needs to have the capacity to accurately identify the hyperpartisan news.

This chapter describes the work presented in the International Workshop on Semantic Evaluation 2019 (SemEval-2019) task 4 Hyperpartisan News Detection task [Kiesel et al., 2019]. Our system[2] uses sentence representations from averaged word embeddings generated from the pre-trained ELMo model with CNN and Batch Normalisation for predicting hyperpartisan news. The final predictions were generated from the averaged predictions of an ensemble of models. In the competition, which had a total of 322 registered teams,

[1] https://pan.webis.de/semeval19/semeval19-web/

[2] The code is available at
`https://github.com/GateNLP/semeval2019-hyperpartisan-bertha-von-suttner`

our system with this architecture ranked in first place among 42 teams who submitted a valid run, based on accuracy, the official scoring metric.

4.1 Problem Definition

Hyperpartisan news is typically defined as news which exhibits an extremely biased opinion in favour of one side, or unreasoning allegiance to one party [Potthast et al., 2017]. The SemEval-2019 Task 4 on "Hyperpartisan News Detection" [Kiesel et al., 2019] is a document-level classification task which requires building a precise and reliable algorithm to automatically discriminate hyperpartisan news from more balanced stories.

4.1.1 Data

Two types of dataset have been made available for this task. The *by-publisher* corpus contains 750K articles which were automatically classified based on a categorisation of the political bias of the news source. This dataset was split into a training set of 600K articles and a validation set of 150K articles, where all the articles in the validation set originated from sources not in the training set. The second set, *by-article*, contains just 645 articles which have been labelled manually. The final evaluation [Potthast et al., 2019] was carried out on a dataset of 628 articles which were also labelled manually. Table 4.1 shows the examples of hyperpartisan news and non-hyperpartisan news related to Donald Trump.

4.1.2 Document Size

One of the major challenges of this task is that the model must have the ability to adapt to a large range of article sizes. For instance, in one of the training data sets, the *by-publisher* corpus, the average article length is 796 tokens, but the longest document has 93,714 tokens. Most state-of-the-art neural network approaches for document classification use a token sequence as network input. For instance, Conneau et al. [2016] padded the input text to a fixed length of 1014, and truncated the text if it was larger than this. Similarly, Zhou et al. [2015] padded the sentence to a fixed length and cut extra words if it was longer than that. However, this implies either a high computational cost when a very large maximum sequence length is used to fully represent the longest articles, or alternatively, potentially a significant loss of information

Reporting style	
Hyperpartisan	*MADISON - U.S. Rep. Mark Pocan (WI-02) released the following statement after the White House announced the Keystone XL Pipeline will be exempt from President Trump ’s executive order requiring construction projects to be built with American steel . ” President Trump has repeatedly vowed to build construction projects with American steel as a way to support manufactures and create jobs here at home — even going as far as making it one of his first signed executive orders , ” Rep. Pocan said . ” But now , he has decided that a major construction project will not abide from his own Buy America plan . House Republicans have already shown their hand by opposing provisions to permanently require American steel be used in our country ’s infrastructure projects . If President Trump is concerned with livelihood of American workers , he should take steps to ensure the Buy America Act is strongly and meaningfully enforced . Unfortunately , the President ’s talk about ‘ Buy America ’ is proving to be just another broken promise to the American people . ” Previously , Rep. Pocan sent a letter to the Trump transition team asking the President to enforce and expand the Buy America Act , and also highlighting Congressional Republican leadership ’s opposition to requirements that American tax dollars go to companies that employ American workers . To date , Rep. Pocan did not receive a response from the transition team , nor have White House officials responded in any capacity . Trump Turns his Back on American Workers*
Non-Hyperpartisan	*US President Donald Trump on Thursday dismissed an upcoming book on his campaign and administration as ” full of lies ” and invented sources , after unsuccessfully attempting to block its release . ” I authorised Zero access to White House (actually turned him down many times) for author of phony book ! I never spoke to him for book . Full of lies , misrepresentations and sources that do n’t exist , ” Mr. Trump tweeted in reference to Michael Wolff ’s ” Fire and Fury : Inside the Trump White House . ” ” Look at this guy ’s past and watch what happens to him and Sloppy Steve ! ” Mr. Trump wrote . It was unclear to whom Mr. Trump was referring , with possibilities including Steve Bannon , his former chief strategist , and Steve Rubin , the president of Henry Holt and Company , which is publishing Wolff ’s book . The book quotes key Mr. Trump aides , including Mr. Bannon , expressing serious doubt about his fitness for office . Mr. Trump has been enraged by the betrayal by Mr. Bannon — a man who engineered the New York real estate mogul ’s link to the nationalist far right and helped create a pro - Trump media ecosystem . After Mr. Trump instructed his lawyers to try to block the release of the book , the publishers responded by moving the release date up by four days to Friday . Trump says new book on his administration ‘ full of lies ’*

Table 4.1: Examples of Hyperpartisan news and Non-Hyperpartisan news in the *by-article* dataset.

if the sequence length is restricted to a manageable number of initial tokens from the d ocument.

To address this, the **E**LMo **S**entence **R**epresentation **C**onvolutional (ESRC) network was developed. This first pre-calculates sentence level embeddings as the average of ELMo [Peters et al., 2018] word embeddings for each sentence, and represents the document as a sequence of such sentence embeddings. It then applies a lightweight CNN, along with Batch Normalization (BN), to learn the document representations and predict the hyperpartisan classification.

Several models were created based on the two datasets, and evaluated using cross-validation on the *by-article* training set (as the final test set was not available to the participants and it was only available for a maximum of three evaluations). A CNN model which used ELMo-based sentence embeddings to represent the article, and was trained on the *by-article* set only, turned out to outperform all other models attempted.

In order to investigate the usefulness of the *by-publisher* training data for training a model that performs well on the manually annotated *by-article* corpus, experiments were performed with various kinds of pre-training and fine-tuning, and it was found that the use of the *by-publisher* corpus was actually harmful and decreased the usefulness of the model. Thus, this work focused on the by-article set. Later, transfer learning is applied to a relatively smaller tendentious news corpus in order to optimise our ability to detect tendentiousness in climate change news.

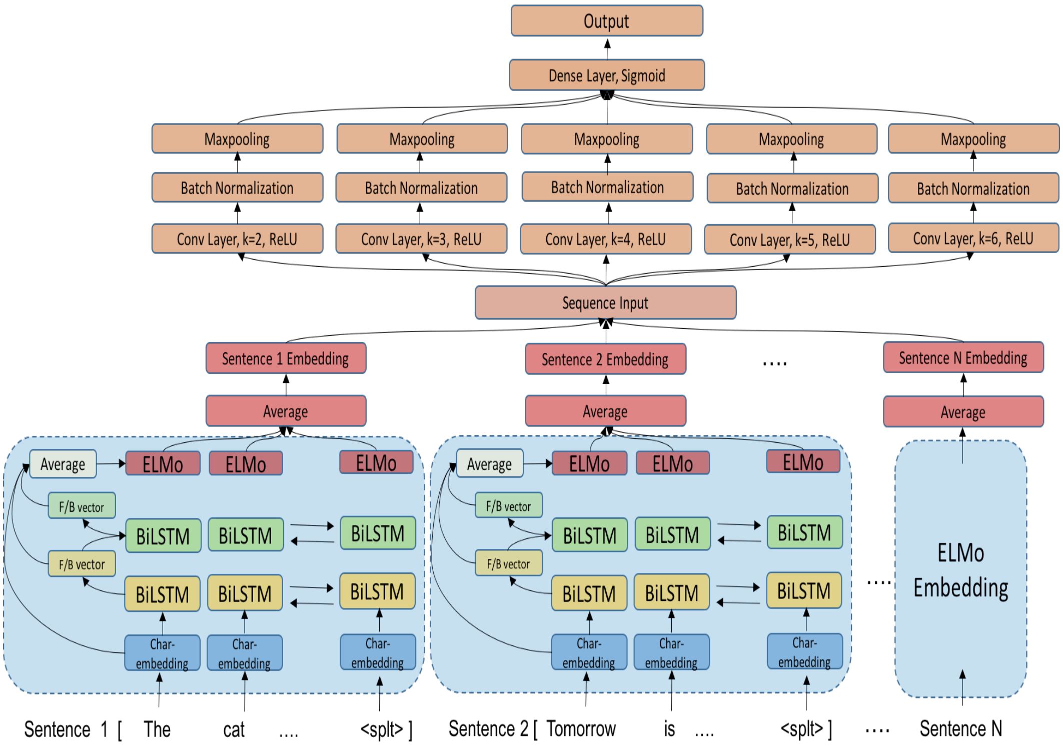


Figure 4.1: System architecture, *F/B vector* denotes Forward/Backward hidden state from BiLSTM layers.

4.2 System Description

This section starts with introducing the ESRC model from bottom to top (see Figure 4.1). It first explains the ELMo embedding layers, and how the word embeddings are extracted and processed to the CNN model. Then, it discusses the architecture of the upper CNN model and its parameters. It also discusses the usefulness of the BN layer in the CNN models.

4.2.1 Deep Contextualized Word Representation: ELMo

Traditionally, the input to a deep learning model is a set of pre-trained word embeddings such as Word2Vec [Mikolov et al., 2013], Glove [Pennington et al., 2014], or FastText [Bojanowski et al., 2017]. In this model, it uses the official AllenNLP[3] library to generate ELMo embeddings, in which the word representation is learned from character-based units as well as contextual information from the news articles. Compared with traditional word embeddings, ELMo produces multiple word embeddings for the same word, depending on the context, which enables the model to distinguish potentially between different meanings of that word. For instance, the term ‘bank’ will have different word embeddings in the context of ‘river bank’ and ‘bank holiday’ respectively, while traditional word embeddings only have a fixed word representation.

Specifically, the first layer of ELMo uses a character-level CNN to transform word strings into a word representation. These character-based word representations allow the model to pick up on morphological features that word-level embeddings could miss, enabling a valid word representation to be formed even for out-of-vocabulary words. For instance, since the term ‘hyper-partisan’ is a relatively uncommon word, most word-level embeddings (e.g., Word2Vec, GloVe, etc.) would be expected to initialise such a term either as random vectors or zeros since the term is not seen in the training, and its vectors therefore cannot be assigned, or are incorrectly assigned, to such terms [Won and Lee, 2018].

To address this, ELMo uses a bi-directional LSTM [Gers et al., 1999] layer to calculate two intermediate word representations separately. In the first bi-directional LSTM layer, the intermediate word vectors are formed through the forward/backward passes that contain information about a certain word and the context before/after that word, from the character-based word representation. Similarly, these intermediate word vectors, from the first bi-directional

[3] https://allennlp.org/

LSTM layer, are fed into the next layer of bi-directional LSTM, which makes it capable of disambiguating the same word into different representations based on its context.

The original[4] pre-trained ELMo model is used to output three representations for each word. Each representation corresponds to a layer output from the ELMo pre-trained model. Since the task submission virtual machine had limited configuration and therefore it was necessary to reduce the computational cost, the average is taken of all three representations to form the final word embedding, and compute the sentence embeddings by averaging the word embeddings in the sentence.

4.2.2 Convolutional Layers

The pre-calculated ELMo sentence embeddings are then fed into the upper CNN model, where it combines five convolutional layers for different filter sizes ($k = 2, 3, 4, 5, 6$), where each has 512 filters. Each convolutional layer is followed by a non-linear activation function ReLU [Nair and Hinton, 2010], which introduces non-linearity after computing linear operations during the convolutional layers. Let $x_i \in \mathbb{R}^d$ be the d-dimensional sentence vector corresponding to the i-th sentence in the document. Each document is padded to contain n sentences, so a hidden state $h_{k,i}$ from convolutional layers is generated by:

$$h_{k,i} = ReLU(W_k \times x_{i:i+k-1} + b_k)$$

where the filter of size k is convoluted from sentence i to *i+k-1* to compute a convolutional weight W_k and a bias b_k. Thus, each filter size k outputs a feature map h_k:

$$h_k = [h_{k,1}, h_{k,2}, ..., h_{k,n-k+1}]$$

4.2.3 Batch Normalization

The feature maps are then fed into Batch Normalization (BN) layers for reducing internal covariate shift in networks [Ioffe and Szegedy, 2015]. Specifically, BN normalizes the input distribution by subtracting the batch mean and dividing by the batch standard deviation, so that the ranges of input distribution between each layer stay similar. This characteristic enables the model to have a higher learning rate and therefore faster training speed. Traditional deep learning models often suffer from overfitting, which refers to a learned feature

[4] `elmo_2x4096_512_2048cnn_2xhighway`

representation that corresponds too closely to a particular dataset, and therefore fails to fit unseen data or predict future observations reliably [Anthony, 2003]. To tackle this, Sun et al. [2017] sparsified the gradient vectors in back propagation by computing a small subset of the full gradient to reduce computational cost and overfitting. Ashiquzzaman et al. [2018] used a dropout approach, which randomly disables a small subset of neurons in the neural network, to prevent neurons from co-adapting too well to the dataset. BN is also used to reduce overfitting [Chang and Chen, 2015, Laurent et al., 2016] by decreasing the dependence of weight initialisation between each layer. The original paper [Ioffe and Szegedy, 2015] suggested that BN should be applied before the activation layer, but the model applies it after the activation layer, after observing better performance in our model this way round. It also applied weighted moving-mean and moving-variance to avoid updating the mean and variance so aggressively in the mini-batch during training time.

Given h_k^j from j-th filter, it has m feature maps in the mini-batch $\mathcal{B} = \{h_{k,1\ldots m}^j\}$. In order to ensure each layer still has optimal weights to transit sentence representations in the network, BN introduces two trainable parameters γ, β to re-scale and re-shift the normalized value y_l, which are defined as follows:

$$\mu\mathcal{B} \leftarrow \frac{1}{m}\sum_{l=1}^{m} h_{k,l}^j$$

$$\sigma^2\mathcal{B} \leftarrow \frac{1}{m}\sum_{l=1}^{m}(h_{k,l}^j - \mu\mathcal{B})$$

$$\hat{h}_{k,l}^j \leftarrow \frac{h_{k,l}^j - \mu\mathcal{B}}{\sqrt{\sigma^2\mathcal{B} + \varepsilon}}$$

$$y_l \leftarrow \gamma\hat{h}_{k,l}^j + \beta$$

BN first calculates the mean $\mu\mathcal{B}$ and variance $\sigma^2\mathcal{B}$ in the mini-batch, and then the internal value $\hat{h}_{k,l}^j$ is transited to a sub-network layer composed of the linear transform to compute the normalized value y_l. BN allows Stochastic Gradient Descent (SGD) or any of its variants to re-scale and re-shift the normalized value by changing only these two parameters γ, β, instead of changing all the weights, which would reduce the stability of the network. The model also applied weighted moving-mean μ_{moving} and moving-variance σ^2_{moving} to avoid updating the mean and variance so aggressively in the mini-batch during training time:

$$\mu_{moving} = \alpha \times \mu_{moving} + (1 - \alpha) \times \mu\mathcal{B}$$

$$\sigma^2_{moving} = \alpha \times \sigma^2_{moving} + (1 - \alpha) \times \sigma^2\mathcal{B}$$

where α is the weight to control how much the mean and variance are updated based on the previous moving statistic. In the ESRC, the α is set to 0.7 as optimal, determined by exploring values from 0.1 to 0.9 at an earlier stage of the experiments.

4.2.4 Max Pooling Layer

Typically, CNN layers are followed by max-pooling [Collobert et al., 2011] layers since this leads to a faster convergence rate by selecting superior invariant features, which improves generalisation performance [Nagi et al., 2011, Kim, 2014]. In the model, it also performs max-pooling operation on the outputs of the batch-normalization layers, and takes the maximum value $\tilde{h}_k = max\{h_k\}$ as the feature corresponding to each filter. The max-pooling operation aims to capture the most salient information (i.e., the one which has the highest value in the feature map).

4.2.5 Dense Layer

Then the outputs of the max-pooling for all convolution layers are concatenated to form the input ($H = (\tilde{h}^1_k, \tilde{h}^2_k, ..., \tilde{h}^j_k)$) to a dense (also called fully connected) layer, which maps to a single output d, followed by the Sigmoid function for the binary classification task:

$$d = Sigmoid(H \circ W + b)$$

where W is the weight matrix and b is a bias term. For configurations, the model used the Adam algorithm [Kingma and Ba, 2014] as the optimizer since it leverages the power of adaptive learning rates methods to find individual learning rates for each parameter. The use of the Adam optimizer was determined after comparing the model performances when using other optimizers.

Since *Sigmoid* activation yield the probability of prediction p where $p \in [0, 1]$ and the true label is y, the model implements binary cross-entropy CE as the loss function as:

$$CE = -(y \log(p) + (1 - y) \log(1 - p))$$

4.3 Experiments

In this task, all models are built using Keras[5] with a Tensorflow backend. All the results are shown in Table 5.2. The table shows for each model the accuracy obtained on the *by-article* training set, and for the submitted models, the *by-publisher* test set, and the hidden *by-article* test set (which unlike the other two, was not available to participants).

4.3.1 Data

The maximum, mean, and minimum numbers of tokens in the *by-article* corpus are: 6470, 666, 19 respectively, and in the *by-publisher* are: 93714, 796, 10 respectively. This makes it impractical to directly use word level representations as the input for our models. For instance, it will result in high computational cost if each article is padded by its maximum number of tokens (i.e., 6470 in the *by-article*, and 93714 in the *by-publisher*), or a significant information loss if each article is truncated with its minimum, or even with the average number of tokens.

As a simple and easy to calculate compromise between representing the details of the article and as much of a longer article as possible, the article is represented as a sequence of sentence embeddings which are calculated as the average of the word embeddings of a sentence. This can be done using any pre-trained word embeddings and does not require a large training set for training or pre-training, so can be easily applied to even the small *by-article* corpus. To form the input sequence for our network, a maximum of the 200 initial tokens per sentence was used for each sentence embedding and a maximum of 200 sentences was used per article (i.e., 40,000 tokens coverage in total). The title of the article was used as the first sentence for each document.

4.3.2 Preprocessing

The model is character-based, taking the morphological information (i.e., the composition of letters) as the input. This enables us to only perform minimal pre-processing. Specifically, it extracts both the title and article text from the original XML representation, since the title could also provide important information to each article [Peramunetilleke and Wong, 2002, Condit et al., 2001].

[5] https://keras.io/

All the original HTML paragraphs in the text cause a sentence break; all text paragraphs have been split into sentences using `spaCy`[6]. The original case of the text was maintained. Whitespace is normalized to a single space between tokens; numbers are replaced by a special number token (i.e., '[NUM]'); and all punctuation and other special characters are preserved as input to the pre-trained ELMo model.

4.3.3 Fine-tuning

In order to investigate the correlation between the two datasets, this study first built the `ESRC-publisher` model which is trained on a randomly selected 100K out of the 750K articles from the *by-publisher* corpus. Since the number of parameters of pre-trained ELMo is so large (93.6 million), it is impractical to generate ELMo embeddings for the entire corpus.

Since the `ESRC-publisher` was trained on a relatively large set, instead of training a new model on *by-article* from scratch, a fine-tuning method is implemented. In practice, training a large deep learning model (i.e., a model with a huge number of parameters) on a small dataset would greatly affect the ability of model generalisation and result in overfitting [Perez and Wang, 2017]. Thus, fine-tuning is a common approach to continue training a large network on a smaller dataset by either truncating the last layer or freezing the weights of the first few layers [Yosinski et al., 2014, Howard and Ruder, 2018].

The former is typically used when the model needs to replace the last prediction layer (e.g., softmax) for a different problem. For instance, in image classification, the last layer of a pre-trained ImageNet has 1,000 categories, but a use-case task has only 10 categories. A new softmax layer with 10 categories will therefore replace the original softmax layer which has 1000 categories. The latter option, freezing the weights of the first few layers, is also a common practice for fine-tuning since the first few layers capture universal features, in order to keep those weights intact. Instead, here the network is forced to focus on learning dataset-specific features in the subsequent layers. For instance, Tajbakhsh et al. [2016] found that the first few layers of CNN can learn low level image features (e.g., pixels), which are applicable to most vision tasks, but the last few layers learn high-level image features (e.g., curves, shapes, etc.) in the vision task.

In this case, since the *by-publisher* has the same categories as *by-article*, the latter approach is used, fine-tuning the `ESRC-publisher` model based on

[6] https://spacy.io/

the *by-article* set to obtain the `ESRC-publisher-article` model by freezing the weights of all but the last layer of the model.

4.3.4 Ensemble Learning

Ensemble learning techniques, which combine the outputs of several classifiers to form an integrated output, have led to the enhancement of classification accuracy [Wang et al., 2011]. Popular ensemble methods, such as Bagging, Boosting and Stacking [Dietterich, 2000], are commonly used in classification tasks.

Specifically, Bagging [Brown and Kuncheva, 2010] uses bootstrap sampling to randomly obtain the data subsets for training the base learners, and the final result is normally calculated by a simple majority vote. Boosting [Dietterich, 2000] uses a weighted version of the training set, in which more weight is given to those data subsets which were misclassified in an earlier round. Stacking [Sollich and Krogh, 1996] combines multiple classification models by averaging the outputs of each individual classifier. In our experiment, the average stacking ensemble is used to ensemble the contribution of each model equally to the final prediction since these models have similar performances, and the predictions could be equally averaged by the ensemble model.

For the evaluation on the hidden test set, the best three models are selected from the 10-folds, according to the accuracy on the evaluation set of each fold to form an average ensemble model, `ESRC-article-BN-Ens`.

4.3.5 Evaluation

The official ranking metric is using accuracy since the target classes in both datasets are balanced (i.e., 50% of hyperpartisan news and 50% of non-hyperpartisan news). The accuracy is calculated by:

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

where $TruePositive$ is the case when the actual class of the article is hyperpartisan and the predicted is also hyperpartisan, $TrueNegative$ is the case when the actual class of the article is not hyperpartisan, and nor is the predicted. $FalsePositive$, however, is where the predicted class is hyperpartisan but the actual class is not, i.e. the model incorrectly predicts the hyperpartisan class.

$FalseNegative$, on the contrary, is where the model predicts the article is not hyperpartisan but the actual class is.

This study trained the `ESRC-article` model only on the *by-article* set, one version with and one version (`ESRC-article-BN`) without the additional batch normalization (BN) layer. The accuracy for the `ESRC-publisher` model is from evaluating on the whole `by-article` training set, while all other evaluations on the `by-article` training set were carried out using a 10-fold cross validation.

The k-fold cross validation [Allen, 1974, Stone, 1974] refers to randomly partitioning the original sample into k equal size subsamples, where a single subsample is retained as the evaluation dataset for testing, and the remaining $k - 1$ subsamples are used as the training set. Then, the cross validation is repeated k times, where each of the k subsamples is used exactly once as the evaluation dataset. However, because of the very limited size of that corpus, the evaluation part of each fold was also used for early stopping (i.e., a method for preventing overfitting by stopping the training epoch before the model has overly learned the training dataset) and model selection within each fold (i.e., select the model with the best performance, which in our case is the highest validation accuracy).

For comparison, the table also shows the results for an earlier version of the model, `GloVe-article`, which used GloVe word embeddings (6 billion words, 300 dimensional) to represent up to the first 400 words of the article, and which did not use batch normalization.

Models	**By-Article Training**
GloVe-article	.7963
ESRC-publisher	.5643
ESRC-publisher-article	.8189
ESRC-article	.8182
ESRC-article-BN	.8387
ESRC-article-BN-Ens	**.8404**
Submitted Models	**By-Article Test**
GloVe-article	.7659
ESRC-article-BN-Ens	.8216
Submitted Models	**By-Publisher Test**
GloVe-article	.6435
ESRC-article-BN-Ens	.5947

Table 4.2: System comparison (accuracy).

4.4 Results and Discussion

The results are shown in Table 5.2. In the experiments, the model uses the *by-article* training set for evaluating model accuracy since the *by-article* test set is only accessible for the final submission in the challenge.

The traditional word embedding GloVe was implemented with the upper CNN model on the *by-article* training set only. The `GloVe-article` yields 79.63% accuracy through 10-fold cross validation. The `GloVe-article` was also submitted to the hidden textitby-article test set, yielding 76.59% accuracy which was the third place among other submissions in the early-bird submission.

Then, this study developed the `GloVe-article` to the ESRC model by implementing ELMo word embedding. It also averages the generated word level embeddings to form sentence embedding in order to cope with the diverse article length. We initially train the ESRC model on the *by-publisher* set and test it on the *by-article* training set (i.e., `ESRC-publisher`). However, the `ESRC-publisher` model performs extremely badly on the *by-article* evaluation data and yields only 56.43% accuracy. This is potentially because the two datasets have fundamentally different features for discriminating partisanship. The *by-article* sets are manually annotated with a pre-defined encoding scheme; the *by-publisher* sets, however, are automatically extracted by a semi-supervised method at the publisher level.

To justify the decision, the `ESRC-publisher` model is fine-tuned by using the *by-article* corpus to produces `ESRC-publisher-article` which performs similarly to a model that is trained only on the *by-article* data (i.e., `ESRC-article`). This confirms results from earlier experiments with simpler models, showing that the use of the *by-publisher* data only hurts the model. The improvement from `ESRC-publisher` to `ESRC-publisher-article` is because the latter overly relies on the features from the *by-article* set.

The performance of `ESRC-article` is enhanced by implementing a BN layer after each convolutional layer. This implementation improves the accuracy by almost 2%. To maximise model performance, it ensembles three best `ESRC-article-BN` models based on the validation accuracy in the 10-fold cross validation. The `ESRC-article-BN-Ens` outperforms others on the *by-article* training set. In the final submission, the `ESRC-article-BN-Ens` was the winning entry.

Overall, the algorithm used for assigning the labels to this dataset just does not reflect information about hyperpartisan articles sufficiently to be helpful.

The `GloVe-article` model also confirms this as its accuracy is even higher than that of the ESRC-article-BN-Ens model on the by-publisher dataset.

A quick manual inspection of the data showed that the source of an article is insufficient by far to identify articles as hyperpartisan or not. For instance, a news with headline "*Daily satellite images of Greenland 's glaciers reveal the break-up of two of its largest glaciers in the last month.*" was labelled as hyperpartisan news in the by-publisher dataset. Since this article was mainly reporting about the melting glaciers in Greenland, it hardly sees any partisanship.

It would be interesting to know how the algorithm used for creating the *by-publisher* corpus actually performs on the *by-article* corpus. To get maximum performance on the *by-article* dataset, it was therefore decided to completely ignore the *by-publisher* data for our final model. The use of BN also showed significant improvement.

4.5 Limitations and Future Work

Since this study uses a CNN with a comparatively large number of parameters in relation to the size of the training set which is rather small, significant variance is expected in the generated models and therefore the average of an ensemble of several models is used for the final predictions. However, the limitation of the average ensemble is that each model has an equal contribution to the final prediction made by the ensemble. Although this is not the case in this experiment, it is problematic when some models are known to perform much better or much worse than others. Future work could implement a weighted ensemble where the contribution of each model to the final prediction is weighted by its performance. For instance, this could be done by using small positive values, which summed together would equal 1, to indicate the percentage of trust or expected performance from each model.

In future work, other architectures and the recent state-of-the-art pre-trained language model Bidirectional Encoder Representations from Transformers (BERT) [Devlin et al., 2018a] could also potentially improve model performance in detecting hyperpartisan news.

4.6 Summary

This chapter presented the details of the work presented in the SemEval-2019 Task 4: Hyperpartisan News Detection. The release of the hyperpartisan news dataset makes it possible to build a deep learning model to automatically detect hyperpartisan news on a large scale. Due to the similarity between hyperpartisan news and tendentious news, this dataset provides an opportunity to implement a transfer learning approach to detect the tendentious news from the knowledge of hyperpartisan news. Participation in this competition was therefore under such motivation, and the model was the winning entry.

The main challenge of this task is that the learning model needs to have the ability of adapting to diverse article lengths. To address this, the ESRC model was developed by averaging ELMo word embeddings to form a sentence representation, and thus each article becomes a sequence of sentence representations that enhance the model adaptability for various document lengths. The experimental result indicates that the implementation of hierarchical document structure, Batch Normalization and contextualized word embedding could significantly improve the model performance.

Although the ESRC model was the wining entry to this competition, a limitation is that the result expects significant variance in the generated model since the CNN model has a large number of parameters, but the size of training samples are relatively small. Also, there are several neural encoders (e.g., LSTM with attention, Transformer) which might potentially outperform the standalone CNN structure.

Chapter 5

Hierarchical Document Representation

The recently released Hyperpartisan News Detection dataset described in the previous section affords great potential for developing methods for the automatic classification of biased news. However, the diversity of document lengths in this dataset produces some challenges. Traditional learning models encode document representation without considering structural information between document and sentence, sentences and words, especially in documents which contain hundreds of sentences, such as newspaper articles. Also, traditional word embeddings also generate a context-free representation which might cause semantic ambiguity.

To address these issues, this chapter extends the work from the previous chapter on hyperpartisan news detection, and proposes a method that combines hierarchical frameworks with recent contextual embeddings to improve the model performance, encoding various sizes of documents effectively. It investigates different neural networks, such as CNN, RNN and Deep Average Network (DAN), by incorporating a hierarchical framework and contextualized word embeddings (i.e., ELMo and BERT). To evaluate this performance, it uses the same dataset and task as in Chapter 4.

5.1 Problem Definition

One of the key issues in text mining and NLP is how to effectively represent documents using numerical vectors [Zhao and Mao, 2018]. Typically, Bag-of-Words (BoW) [Lan et al., 2009, Joachims, 1998], Latent Dirichlet Allocation (LDA) [Blei et al., 2003] and n-grams with Term Frequency-Inverse Document

Frequency (TF-IDF) [Wu et al., 2008] are used to generate document representation. In the BoW model, documents are represented as the 'bag' of words, disregarding grammar and word order, where each element in a BoW vector representation denotes the normalized number of occurrences of a term in the document. However, this method simply conducts exact word matching (also called hard mapping) to count the frequency of a term, and might cause sparsity of document representation and dimensional explosion [Zhao and Mao, 2018]. Moreover, term frequency is not the best representation for the text, since common words (e.g., stop-words, such as 'the', 'a' and 'to', etc.) almost always have the highest frequency in the text. Having a high raw count therefore does not mean that the corresponding word is more important than other words. To address this, TF-IDF representation starts with calculating the term frequency (TF) in the documents, and then uses inverse document frequency (IDF) factor to diminish the weight of terms that occur very frequently in the document set and increase the weight of terms that occur rarely [Jones, 1972]. However, it also computes document representation based on word-count space, which may be slow for large vocabularies, and makes no use of semantic similarities between words [Kim et al., 2019]. finally, although LDA could generate a dense document representation while capturing semantic relations in the text [Rafi et al., 2011], this assumes a fixed vocabulary of word types, and makes it hard to handle out-of-vocabulary (OOV) issues in unseen documents [Das et al., 2015]. For instance, given a new document, since there are unknown words occurring, LDA therefore would need to be retrained to get the representation for the new document.

Recently, neural network models, which incorporate low-dimensional word embeddings, outperform most feature engineering-based models. For instance, Zhang et al. [2015] developed a character level CNN for text classification. Their CNN model outperforms other models which are based on BoW, TFIDF and their variants, on several benchmark datasets, such as Yelp reviews, the DBpedia ontology, and Amazon reviews. Ma et al. [2016] developed an RNN model that learns the embedding matrix from scratch to detect rumours from microblogs, and outperforms SVM based models. Ruchansky et al. [2017] utilised doc2vec [Le and Mikolov, 2014], which is an extended version of word2vec, for generating document representations to detect fake news. Their model also outperforms traditional feature engineering based models. However, such neural network models imply either the maximum sequence length is used to fully represent the longest document, which causes a high computational cost, or alternatively a significant information loss if the sequence length is restricted to a manageable number of initial tokens from the documents. Furthermore, this kind of document representation ignores the hierarchical features of a document, such as the structural relationship between word and sentence, or between sentence and document.

In an attempt to resolve this issue, this chapter extends the work of the ESRC model, and proposes a hierarchical framework that captures structural features between word and sentence, and between sentence and document, and also utilises contextualised document representation. Traditionally, the input to a neural network model is a set of pre-trained word embeddings such as Word2Vec [Mikolov et al., 2013], Glove [Pennington et al., 2014], or FastText [Bojanowski et al., 2017]. Such word embeddings generate a context-free representation for each word in the vocabulary. For instance, the word "apple" has the same vector representation for the meanings "eat apple" and "the Apple company"; however, the context completely changes the meaning of "apple" in a sentence. For this reason, recent pre-trained language models (e.g. ELMo [Peters et al., 2018] and BERT [Devlin et al., 2018b]) utilise bidirectional approaches to guard against context-free issues. Specifically, ELMo and BERT respectively use bidirectional LSTM and Transformer Vaswani et al. [2017b] to learn the contextual information from the text. ELMo consists of three layers, comprising a character CNN learned from character-based units to pick up on morphological features, and two bidirectional LSTM layers for capturing contextual information before and after each word [Peters et al., 2018]. BERT [Devlin et al., 2018b], on the other hand, is based purely on attention mechanism [Vaswani et al., 2017a], which learns contextual relations between words in a text by encoding the left and right context of each word in the sentence.

In this model, each sentence is represented by implementing a specific neural network architecture to encode the contextual word embeddings in the sentence. Similarly, the document is represented by encoding all sentence representations which are generated from the previous step. In order to evaluate the hierarchical contextual document representation, this chapter implements a document-level classification task based on a publicly accessible dataset, and compares the performances between models. It also compares the contextual word embeddings ELMo and BERT, in different types of hierarchical frameworks.

5.2 Related Work

Traditionally, many feature engineering based approaches have been used to classify documents. For instance, Rubin et al. [2016] used TF-IDF representation with SVM to classify satirical news articles. Similarly Fortuna et al. [2009] also represented news articles in the vector space model by using TF-IDF weighting, and utilised SVM to identify the bias in describing the events in news articles. However, such feature engineering-based models suffer from sparsity of document representation and dimensional explosion.

Recently, neural network approaches have been used to generate document representations and outperform many feature engineering-based methods. For instance, Iyyer et al. [2014] applied a recursive neural network to identify political ideology evinced by sentence-level representation, and the network outperforms logistic regression with BoW models. Kim [2014] pioneered CNN on document classification, with the CNN model achieving highest accuracy on an IMDB dataset against other models. However, such approaches generate document representations without considering the characteristics of the document structure hierarchically.

To address this issue, Yang et al. [2016a] developed a Hierarchical Attention Network (HAN), which could capture the hierarchical features on both sentence level and document level through a stacked RNN architecture. In their implementation, the HAN assumes that each sentence is composed of words with a different level of importance to that sentence; similarly, each document is composed of sentences with different levels of importance. Thus, they utilised the attention mechanism to capture the importance on both the sentence level and document level. Such implementation outperformed many classifiers and indicates that such prior hierarchical information has the potential to generate better document representations, especially when the document sizes are in a wide range. For instance, Zheng et al. [2019] compared different hierarchical encoders on documents of different lengths, and found that hierarchical frameworks outperform the corresponding neural network models without the hierarchical architecture for document classification. They also indicated that the benefits resulting from the hierarchical architecture can be strengthened as the document length increases.

Generally, hierarchical models have been implemented for many NLP downstream tasks, such as text generation and text classification. For instance, Li et al. [2015] implemented a hierarchical auto-encoder on both word and sentence level, and decoded each representation to reconstruct the original paragraph. Gao et al. [2018] constructed a hierarchical convolutional attention model that utilised a combination of self-attention and target-attention. Abreu et al. [2019] combined RNN with CNN in a hybrid hierarchical attentional neural network in the document classification task.

On the other hand, the input to a neural network model is typically a set of pre-trained word embeddings such as Word2Vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014] or FastText [Bojanowski et al., 2017]. Such word embeddings generate a context-free representation for each word in the vocabulary. To address this issue, context-sensitive word embeddings have recently been developed, such as Embeddings from Language Models (ELMo)[Peters et al., 2018] and Bidirectional Encoder Representations from Transformer (BERT) [Devlin et al., 2018b], which generate a representation based on its context in the sentence by using Bidirectional LSTM [Gers et al., 1999] and Transformer [Vaswani et al., 2017b] respectively, achieving state-of-the-art performance in many downstream NLP tasks, such as machine translation [Zhu et al., 2020], question answering [Devlin et al., 2018b], and text generation [Zhang et al., 2019].

The previous chapter introduced an ESRC model which combined ELMo embeddings with a lightweight CNN model [Jiang et al., 2019b], and the resulting system was ranked first in the SemEval 2019 task 4. Alsentzer et al. [2019] also implemented BERT embeddings on a clinical corpus. Wiedemann et al. [2019] explored the word sense disambiguation of three contextual word embeddings (BERT, ELMo and Flair [Akbik et al., 2018]), and demonstrated that the pre-trained BERT model was able to place polysemic words into distinct 'sense' regions of the embedding space. This chapter extends the utilities of contextual word embeddings by incorporating document structural information, and also compares the performance of ELMo embeddings and BERT embeddings in terms of the document classification task.

5.3 Methodology

Hierarchical frameworks utilise the document structural features such as the relation between word and sentence, and between sentence and document. In order to investigate the effectiveness of the learning model encoding document representation hierarchically, this section compares the different neural network structures with/without incorporating hierarchical frameworks. To evaluate them, this study first establishes three different network structures, without considering structural features as baseline models. Then, it applies hierarchical structures as hierarchical models accordingly on the top of these baseline models. Two different contextual embeddings (ELMo and BERT) are used in both baseline and hierarchical models.

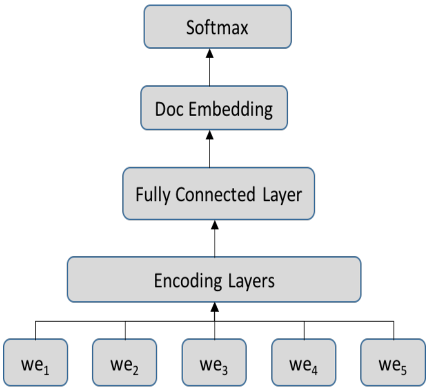


Figure 5.1: Baseline model structure

5.3.1 Baseline Models

Three different network structures are implemented as the baseline models. Figure 6.1 demonstrates the overall baseline model structure. Formally, each document representation is generated from the initial tokens in the document. This is an aggregation of all the contextual word embeddings we from a specific neural network in the encoding layer. Finally, a Fully Connected (FC) layer with softmax activation and Adam optimizer is made for the final classification.

CNN-base: For a possible variant CNN structure, a lightweight CNN model is implemented based on the ESRC. It consists of 128 filters and 7 different convolutional filter sizes [1,2,3,4,5,6,7] with ReLU activation, followed by a batch normalization layer and a max-pooling layer. The results from max-pooling layers are concatenated and go through an FC layer with 32 hidden units and ReLU non-linearity. The convolutional layers therefore can be denoted as follows:

RNN-base: Bidirectional LSTM (Bi-LSTM) is applied as the baseline RNN model. Bidirectional RNN concatenates both forward and backward hidden states, and this characteristic could capture contextual information in each sequence. The Bi-LSTM layer has 100 dimensional hidden units with a dropout probability of 0.2, and is followed by an FC layer with 32 hidden units and ReLU non-linearity.

DAN-base: DAN [Iyyer et al., 2015] implements neural bag-of-words functions that ignore the sequence order information, but significantly increase training speed and could also achieve comparable model performance. It directly takes the average of a fixed number of initial word embeddings to form the document representation, followed by an FC layer with 32 units and ReLU non-linearity. It also uses dropout function before contextual embeddings pass to the encoding layer.

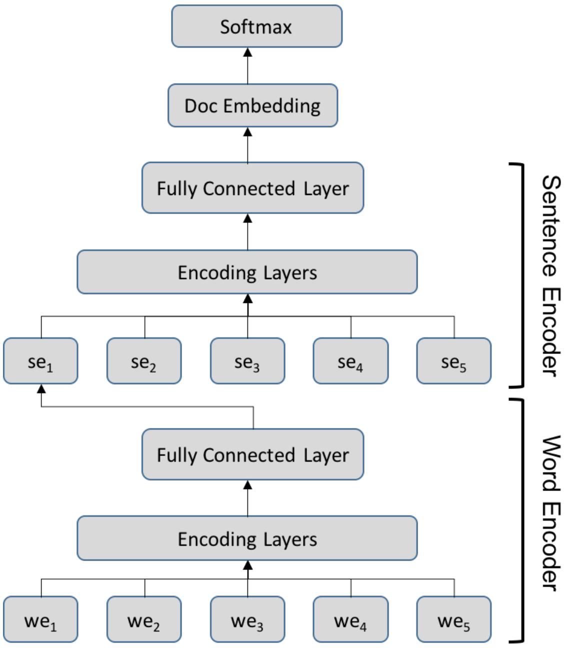


Figure 5.2: Hierarchical model structure

5.3.2 Hierarchical Models

For these, this study utilises the hierarchical features on top of our baseline models. Figure 6.2 demonstrates the overall hierarchical framework structure. The hierarchical models take word and sentence representation as inputs separately. The contextual word embeddings *we* are aggregated to a sentence representation using a specific hierarchical neural network. The document representation can then be formed by aggregating all the sentence representations *se*. Finally, an FC layer with softmax activation and Adam optimizer is made for the final classification.

Let d denote a document consisting of a sequence of sentences (s_1, s_2 , ... , s_m); Meanwhile, let s_i denote a sentence consisting of words ($w^1_{s_i}$, $w^2_{s_i}$,..., $w^n_{s_i}$) where $i \in [1, m]$, the model embeds s_i into a distributional space x = (x_1, x_2, ... , x_n) where $x_j \in \mathbb{R}^k$, $j \in [1, n]$ and k is the dimension of the n-th word embedding in the sentence. All the models are trained to minimise the cross-entropy error by:

$$\ell(\tilde{y}) = \sum_{p=1}^{b} y_p \log(\tilde{y}_p) \tag{5.1}$$

where $y, \tilde{y}$ are the ground-truth label and predicted label respectively, b denotes number of classes.

H-CNN: In the H-CNN model, the word encoder has 128 filters and 7 different convolutional filter sizes $h \in$ [1,2,3,4,5,6,7] with ReLU activation, with each convolutional layer followed by a batch normalization and a max-pooling layer. The results from the max-pooling layers are concatenated to form a sentence representation. The sentence encoder takes each sentence representation as input, with the same structure as the word encoder, except it has an extra FC layer with 32 hidden units and ReLU non-linearity after the final concatenation. Specifically, the convolutional layer using different filter operators $W_{h,j} \in \mathbb{R}^{h \times k}$ is applied to a window of h words to produce a new feature c_j^h at the word level:

$$c_{x_j}^h = BN\left(ReLU\left(x_{j:j+h-1} \circ W_{h,j} + b_{h,j}\right)\right) \tag{5.2}$$

where the notation $\circ$ denotes element-wise multiplication, $ReLU$ denotes the nonlinear function, $b_{h,j} \in \mathbb{R}$ is a bias term, BN denotes batch normalisation.

Then, the max-over-time pooling function is used to capture the most important feature $\tilde{c}_{x_j}^h$:

$$\tilde{c}_{x_j}^h = Max\left(c_{x_j}^h\right) \tag{5.3}$$

The final feature maps c_{x_j} are formed by concatenating all $c_{x_j} = (\tilde{c}_{x_j}^1, \tilde{c}_{x_j}^2, ..., \tilde{c}_{x_j}^7)$, then the sentence representation s_i can be generated by an FC layer:

$$s_i = ReLU\left(c_{x_j} \circ W_j + b_j\right) \tag{5.4}$$

where W_j is a weight matrix and b_j is a bias term. Then, the final document representation d can be obtained similarly: it first obtains the feature maps c_i^h by convoluting the sentence sequence using different filter operators, and applying batch normalisation:

$$c_{s_i}^h = \left(c_{s_1}^h, c_{s_2}^h, ..., c_{s_i:s_{i+h-1}}^h\right) \tag{5.5}$$

Then, the max pooled features can be obtained:

$$\tilde{c}_{s_i}^h = Max\left(c_{s_i}^h\right) \tag{5.6}$$

Finally, after concatenating $\tilde{c}^{h}_{s_i}$ to obtain c_{s_i} the document representation d can be formed as d:

d = ReLU$(c_{s_i} \circ W_i + b_i)$ (5.7)where W_i is a weight matrix and b_i is a bias term, $ReLU$ is the non-linear function. Finally, the document representation d is formed to make the final prediction in a softmax layer.

H-RNN: Two Bi-LSTM encoders are applied to form the H-RNN model. The word-encoder Bi-LSTM has 100 dimensional hidden units with a dropout probability of 0.2, and is followed by batch normalization and an FC layer with 100 hidden units and ReLU activation. The sentence encoder also has the same structure as the word encoder, except it has an extra FC layer with 32 hidden units and ReLU non-linearity.

Formally, the forward $\overrightarrow{r_{x_n}}$ and backward $\overleftarrow{r_{x_n}}$ hidden states at the word level can be obtained by using bidirectional LSTM:

$$\overrightarrow{r_{x_n}} = \overrightarrow{LSTM}\,(x_{1:n}) \tag{5.8}$$

$$\overleftarrow{r_{x_n}} = \overleftarrow{LSTM}\,(x_{1:n}) \tag{5.9}$$

Then the $\overrightarrow{r_{x_n}}$ and $\overleftarrow{r_{x_n}}$ can be concatenated as $r_{x_n} = (\overrightarrow{r_{x_n}}; \overleftarrow{r_{x_n}})$and pass to ReLU non-linear function to form the sentence representation s_m:

$$s_m = ReLU\,(r_{x_n} \circ W_n + b_n) \tag{5.10}$$

where W_n denotes a weight matrix and b_n denotes a bias term. Similarly, the sentence level hidden states can also be formed by:

$$\overrightarrow{r_{s_m}} = \overrightarrow{LSTM}\,(s_{1:m}) \tag{5.11}$$

$$\overleftarrow{r_{s_m}} = \overleftarrow{LSTM}\,(s_{1:m}) \tag{5.12}$$

Then the $\overrightarrow{r_{s_m}}$ and $\overleftarrow{r_{s_m}}$ can be concatenated as $r_{s_m} = (\overrightarrow{r_{s_m}}; \overleftarrow{r_{s_m}})$ and pass to ReLU non-linear function to form the document representation d:

$$d = ReLU\,(r_{s_m} \circ W_m + b_m) \tag{5.13}$$

where W_m denotes a weight matrix and b_m denotes a bias term. Finally, the document representation d is formed to make the final prediction by a softmax layer.

H-DAN: Similar to other hierarchical models, the word encoder takes word embeddings as the input, and then takes the average of the word level representation to form the sentence representation. The sentence encoder then takes the average of the sentence representations to form the document representation. Finally, the document representation is passed through an FC layer that contains 32 units and ReLU non-linearity.

Specifically, the averaged sentence embedding $\tilde{s}$ can be obtained simply by averaging each word embedding x_j in a sentence:

$$\tilde{s} = 1/|n| \sum_{j=1}^{n} x_j^l \quad (l \in [1, k] \text{ and } \tilde{s}, x_j \in \mathbb{R}^k) \tag{5.14}$$

Then, it is passed to a non-linear function *ReLU* to obtain the final representation s_i:

$$s_i = ReLU(\tilde{s} \circ W_j + b_j) \tag{5.15}$$

where s_i denotes the sentence representation, W_j denotes a weight matrix and b_j denotes a bias term. Similarly, the document representation can be formed by:

$$\tilde{d} = 1/|i| \sum_{m=1}^{i} s_i^l \quad (l \in [1, k] \text{ and } \tilde{d}, s_i \in \mathbb{R}^k) \tag{5.16}$$

Then, it is passed to a non-linear function *ReLU* to obtain the final representation:

$$d = f(\tilde{d} \circ W_i + b_i) \tag{5.17}$$

where W_i denotes a weight matrix and b_i denotes a bias term. Finally, the document representation is passed through an FC layer that contains 32 units and ReLU non-linearity.

5.3.3 Embedding Generation

The pre-trained BERT[1] and ELMo[2] models are used to generate contextual word embeddings for baseline and hierarchical models. For generating ELMo embeddings, the official pre-trained ELMo model from AllenNLP is used. The original ELMo pre-trained model generates three vectors for each word, where each vector corresponds to a layer output from the ELMo pre-trained language model. The first layer corresponds to the context-insensitive token representation, followed by the two LSTM layers. Then, the average is taken of all

[1] *BERT-Base, Uncased*

[2] `elmo_2x4096_512_2048cnn_2xhighway`

three vectors to form the final word vector. Specifically, the word representation is learned from character-based units as well as contextual information from the news articles. These character-based word representations allow it to pick up on morphological features that word-level embeddings could miss, and a valid word representation can be formed even for out-of-vocabulary words. Furthermore, ELMo uses two bi-directional LSTM [Gers et al., 1999] layers to learn the contextual information from the text, which makes it capable of disambiguating the same word into different representations based on its context.

This study implements bert-as-service [Xiao, 2018] to generate BERT embeddings. The BERT-base model [Devlin et al., 2018b] contains an encoder with 12 Transformer blocks, 12 self-attention heads and the hidden size of 768. For text classification tasks, BERT normally takes the final hidden state of the special token [CLS] as the representation of the whole sentence [Sun et al., 2019]. Since it needs to generate word-level embeddings, this study implements 'NONE' pooling strategy in the bert-as-service. This implementation generates a fixed size of 768 dimensional vectors for each word in a sequence.

5.4 Experiments

The dataset is split into training and test sets with a ratio of 9:1 and perform 10-fold cross validation on the training set. The final scores are obtained based on the average of 5 predictions on the test set.

5.4.1 Dataset

The Hyperpartisan News Detection dataset[3] contains two parts, as described in the previous chapter. The *By-Publisher* corpus contains 750K articles which were automatically classified, based on a categorisation of the political bias of the news provider. The *By-Articles* corpus contains 1,273 articles which were annotated manually. Although the *By-Publisher* corpus has great potential in training deep learning models due to its significant size, last chapter already revealed that there is no significant correlation between the two corpora, in the sense that training a learning model on the *By-Publisher* corpus leads to low performance in the task of predicting partisanship on the *By-Article* corpus. Thus, in this chapter all models are only trained on the *By-Article* corpus,

[3] *https://pan.webis.de/semeval19/semeval19-web*

Dataset	Hyperpartisan By-Article set
No. of classes	2
No. of documents	645
No. of average sentences/document	31.17
No. of maximum sentences in document	257
No. of average words/sentence	121.13
No. of maximum words in document	5906
No. of average words/document	615.99
No. of words in vocabulary	26135

Table 5.1: Statistics of dataset

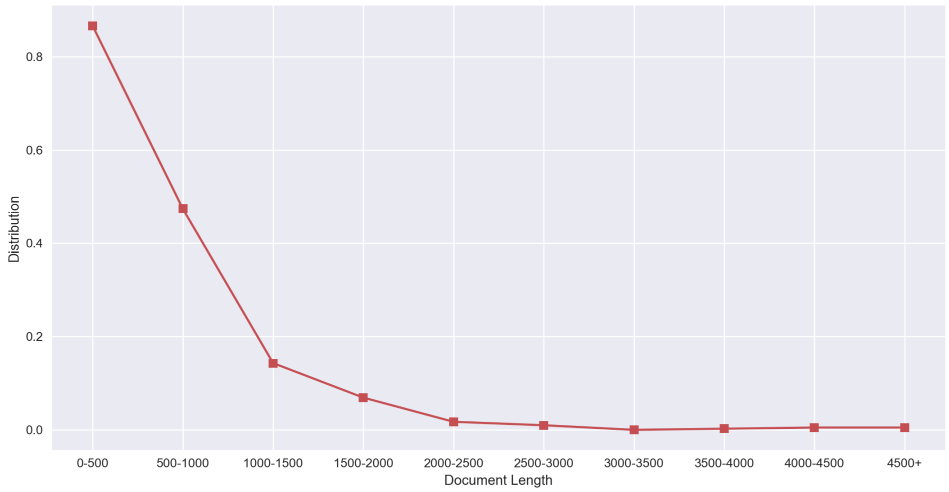


Figure 5.3: Document length distribution

as this is more reliable based on its manual annotation assessment Vincent and Mestre [2018a], and it is also the official ranking corpus for the task. The training set (645 articles) of the *By-Article* corpus was only used, as the rest (628 articles) of the corpus is unavailable to the public (only used for system evaluation).

The statistics of `By-Article` is also calculated as shown in Table 5.1, and the document length distribution as shown in Figure 5.3.

As discussed previously, such a large differentiation in document size makes it impractical to directly use word level representations as the input, as most news articles have no limitation on sequence length compared to other types of sources (e.g. reviews, tweets, etc). In order to calculate the compromise between representing a summary of the article and as much of its full content as possible, this study uses the initial 512 (i.e. the maximum sequence length which can be taken from the pre-trained BERT model) tokens to represent each article in the baseline models. For the hierarchical models, it takes a maximum of 100 words per sentence, and 30 sentences per document.

5.4.2 Pre-processing

The title and article text are extracted from the original XML file, and represent each article as a sequence of sentences. The text paragraphs are split into sentences and white spaces are normalised, as before. As the ELMo pre-trained model is character-based, this enables us to only perform minimal pre-processing. In terms of generating BERT embeddings, the original text is in lower case, and the punctuation is removed.

5.4.3 Evaluation

In this experiment, different metrics for comprehensively evaluating the model performances are compared. Beside accuracy, the precision, recall and F1 are also calculated as follows:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5.18}$$

where precision is used to measure how well the model detects positive values. High precision is important when the costs of false positives are high. For instance, lower precision might mean an email spam detection system falsely identifies important emails as spam.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5.19}$$

where recall calculates how many actual positives that model can capture. This is important when the costs of false negatives are high. For instance, the consequence can be very bad for the bank if a fraudulent transaction is detected as non-fraudulent.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.20}$$

In terms of F1 score, this is the weighted average of Precision and Recall. A good F1 score means that the model has low false positives and low false negatives. It is considered perfect when it reaches 1, while the model is a total failure when F1 reaches 0.

5.5 Results and Discussion

The results presented in Table 5.2 show that, on average, the hierarchical models outperform the baseline models in terms of accuracy. Specifically, the baseline models have difficulty handling a wide range of document lengths, especially if document sequences are truncated which could potentially cause information loss. Accordingly, the use of hierarchical models, which summarise the importance both on the word level and sentence level features by the corresponding encoders, leads to an improvement in accuracy.

Models	**Accuracy (std.)**	**Precision (std.)**	**Recall (std.)**	**F1 (std.)**
RNN-base-ELMo	.7785 (.0505)	.8082 (.0385)	.5919 (.0865)	.6833 (.0710)
RNN-base-BERT	.7692 (.0312)	.7731 (.0281)	.6219 (.0789)	.6893 (.0601)
CNN-base-ELMo	.7704 (.0323)	.7846 (.0592)	.6137 (.0577)	.6887 (.0387)
CNN-base-BERT	.7871 (.0358)	.7946 (.0611)	.5937 (.0549)	.6796 (.0379)
DAN-base-ELMo	.7798 (.0421)	.7831 (.0212)	.6134 (.0800)	.6879 (.0381)
DAN-base-BERT	.7898 (.0371)	.7979 (.0303)	.6040 (.0873)	.6875 (.0571)
H-CNN-ELMo	.8189 (.0471)	.7029 (.0281)	.7857 (.1024)	.7420 (.0895)
H-CNN-BERT	.8334 (.0538)	.7320 (.0329)	.7657 (.1480)	.7485 (.0941)
H-RNN-ELMo	.8091 (.0987)	.7534 (.2193)	.7762 (.0461)	.7646 (.0783)
H-RNN-BERT	.8119 (.1292)	.7743 (.3513)	**.8262 (.3361)**	**.7994** (.3633)
H-DAN-ELMo	.8315 (.0992)	**.8401 (.0031)**	.7239(.0913)	.7776(.0531)
H-DAN-BERT	**.8450 (.0482)**	.8338 (.0431)	.7677(.1096)	**.7993(.0682)**

Table 5.2: Performance comparison between models. Best values are marked in **bold**, standard deviations in parentheses

Interestingly, the accuracy of DAN models (i.e. DAN-base and H-DAN) is higher than that of other neural network structures (i.e. RNN-base, CNN-base, H-RNN and H-CNN). This indicates that simply taking the average is better than other architectures (i.e. RNN, CNN). This might be because the data set is overfitting in either the RNN or CNN models, since they havea larger number of parameters in the convolutional and LSTM layers. Although DAN utilises the unordered functions, the positional information is kept in the word representation when ELMo and BERT are generating word level embeddings based on the context of the word. However, such neural network architectures still have the potential to outperform methods which take the average by hyperparameter tuning and adding training samples.

The accuracy of H-RNN, H-CNN and H-DAN shows around 5% improvement compared to the RNN-base, CNN-base and DAN-base respectively. However, the training speed of DAN is much faster than others in both base and hierarchical structures, while obtaining comparable performance.

Additionally, all the baseline models achieve significantly lower recall scores than hierarchical models. This is because the baseline models miss matching

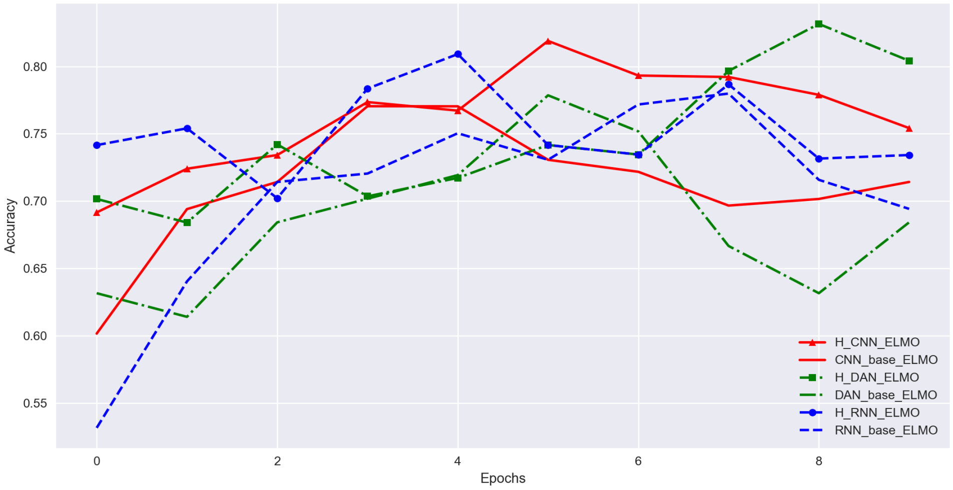


Figure 5.4: Classification accuracy of the training models with ELMo embeddings

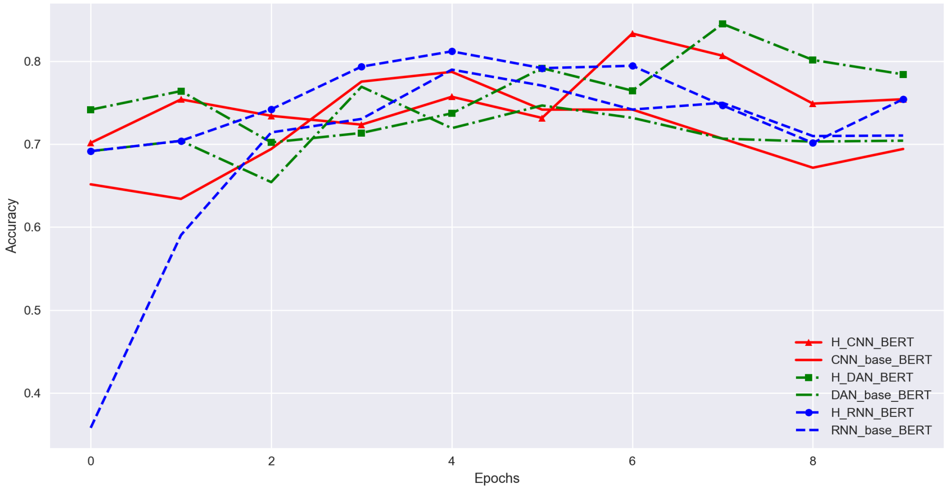


Figure 5.5: Classification accuracy of the training models with BERT embeddings

instances in the training, as the truncated sequence length cannot fully represent features in the document. Also, BERT models are relatively better than ELMo models on average in terms of accuracy. The precision of RNN models are generally higher than CNN models especially in the baselines, but the DAN models still have better precision compared with CNNs and RNNs. However, the DAN models do not yield the highest F1 scores but rather the H-RNN-BERT. This might be because the RNN model consider the position information of each token in the sentence and also each sentence in the document, but the DAN does not.

Figures 5.4 and 5.5 demonstrate the converging curves of ELMo and BERT models respectively. Generally, most of the models converged in the first 10 epochs, due to the large number of parameters and relatively small size of the dataset. Particularly, the baseline models converge more quickly than the hierarchical models generally, as the baselines overfitted quickly. For instance, the accuracy reaches its peak at around the fourth epoch in the baseline models, while the accuracy of hierarchical models mostly converges after the sixth epoch.

5.6 Summary

This chapter has explored the performance of hierarchical models with context-sensitive embeddings on the recently introduced Hyperpartisan News Detection dataset. Specifically, it first compared the hierarchical framework with the baseline models. The results demonstrate that the hierarchical model has the advantage of handling various document sequences and reducing information loss by incorporating structural features in the document. The ELMo and BERT embeddings are also compared in both baseline and hierarchical structures. The results indicate that BERT embeddings generate a better document representation than ELMo in terms of model accuracy in this task. Meanwhile, the DAN models outperform others in generating document representation. In conclusion, the combination of hierarchical frameworks and contextual embeddings could significantly improve model performance in document classification.

Since the hierarchical model is able to encode various document lengths and also generates effective document representation, future work will take account of other additive feature representations (e.g., document-topic distribution and topic-term distribution from the LDA model) to enhance the encoding ability of the hierarchical model on both word level and sentence level. Also, the method could be used for other text classification tasks when there are various document lengths.

Chapter 6

Hierarchical Topic-Aware Neural Networks

In the previous chapter, it investigated the effectiveness of the implementation of hierarchical frameworks combined with different contextualized word embeddings. Typically, the performance of machine learning models can be also enhanced by adding feature representation. For instance, Founta et al. [2019] implemented a unified model which takes text data and metadata (i.e., account age, location, etc.) separately as inputs for detecting abusive tweets. Experimental results demonstrated that combining different model inputs could improve the model accuracy on classification tasks.

Thus, based on the hierarchical frameworks developed in the previous chapter, this chapter looks at the distributions generated from Latent Dirichlet Allocation (LDA) topic models, which enable us to enrich the feature space by adding word co-occurrence distribution and local topic probability in each document. In this chapter, the LDA distributions are regarded as additive features on the sentence level and document level respectively. Second, it compares the performance of different popular neural network architectures incorporating these LDA distributions on the hyperpartisan newspaper article detection task described in the previous two chapters.

6.1 Problem Definition

Traditional methods such as Latent Semantic Analysis (LSA) [Deerwester et al., 1990], probabilistic Latent Semantic Analysis (pLSA)[Hofmann, 1999] and Latent Dirichlet Allocation (LDA) [Blei et al., 2003] have been implemented to infer the semantic meaning of documents through a set of topic representations. Such methods convert text into vector representations which make it feasible for machines to "understand" the semantics of text for many NLP tasks. For instance, Gong and Liu [2001] used LSA to identify semantically important sentences for creating document summaries. They decomposed a document into individual sentences for creating a document representation via a sentence matrix, and performed the singular value decomposition (SVD) on it to reduce its dimensionalites whilst deriving the latent semantic structure from sentence matrix. Brants et al. [2002] claimed that use of pLSA enables the effective representation of sparse information in a text block to be generated. pLSA was used to compare the distance between two blocks of text and select segmentation points based on the similarity values between pairs of adjacent blocks. Kim et al. [2019] implemented document-topic distribution that was generated from LDA, as the document representation for a classification task.

Recently, neural network-based models, which have been proposed in order to generate low-dimensional vector representations, and which are also able to capture semantic word relationships, have been found to outperform most BoW-based models [Ma et al., 2016, Wei et al., 2016]. For instance, the Continuous Bag of Words (C-BoW) model [Mikolov et al., 2013] encodes each word into a fixed length vector representation based on other words surrounding the target word. However, such models suffer from the disadvantage that they do not utilise the word co-occurrence of the entire corpus. Specifically, they only scan the textual information within a local context window, which fails to make use of statistical information of the whole corpus. GloVe [Pennington et al., 2014] attempts to resolve this by implementing both global matrix factorisation and local content window-based methods; however, our proposal uses a different approach that combines the global co-occurrence information with semantic features of local content windows. Another problem is that many neural network models [Yin and Schütze, 2016, Conneau et al., 2017] ignore the hierarchical features of a document, such as the structural relationship between word and sentence, or sentence and document. In an attempt to resolve these issues, based on the hierarchical frameworks that developed in the previous chapter, this chapter extends the hierarchical frameworks by incorporating LDA distributions on word level and sentence level separately.

In order to evaluate the topic-aware hierarchical document representation, this chapter implements a document classification task based on the Hyperpartisan News Detection Task. The documents in this corpus are by nature more challenging for learning models than those typically used for traditional document classification (e.g., IMDB, Amazon reviews). For instance, the document size, which has been discussed in the previous chapter, made it difficult for learning models to encode diverse sequence lengths. Also, unlike sentiment, which typically has an explicit expression or narrative in the text, partisanship is expressed in a more implicit and subtle way in the news articles.

This study performs an evaluation by comparing different popular neural network architectures, with and without incorporating LDA-based distributions, and also compares these with non-hierarchical structures. The code of the model LDA-HAN[1] is available for replicability. Theoretically, the models incorporating LDA distributions should enrich the feature space by adding co-occurrence statistics features and local topic probability distribution on the word and sentence level respectively. Our experimental results demonstrate that the topic-aware document representation outperforms traditional ones, and also that the inclusion of the LDA features has greater impact on the hierarchical representations.

6.2 Related Work

Traditionally, BoW-based approaches have often been used to classify newspaper articles. For instance, Rubin et al. [2016] used a BoW representation with a Support Vector Machine (SVM) to classify satirical news articles. Fortuna et al. [2009] also represented news articles in the vector space model by using TF-IDF weighting, and utilized SVM to identify the bias in describing events in news articles, while Budak et al. [2016] used SVM to quantify news bias in a large set of political articles. Meanwhile, LDA has been combined with traditional feature engineering-based methods in many document classification tasks. Wu et al. [2015] combined LDA with SVM to classify Chinese news, outperforming the models which generate high-dimensional feature space such as TF-IDF models. Li et al. [2016] implemented LDA with a softmax regression to overcome the high dimensional problems of the news text. Kim et al. [2019] regarded the document-topic distribution from LDA as a document representation in which both word frequencies and semantic information are considered, to enhance the performance of document classifiers.

Recently, neural network approaches have been combined with LDA for generating document representations. For instance, Liu et al. [2015] applied

[1] https://github.com/yjiang18/LDA-HAN

LDA to build topic-based word embeddings based on both words and their topics. They implemented LDA with Gibbs sampling [Griffiths and Steyvers, 2004] to assign latent topics for each word token, and then the topic-word distribution can be used to learn topic word embeddings. Xu et al. [2016] also implemented LDA to capture topic-based word relationships and then integrated it into distributed word embeddings. Wang and Xu [2018] implemented LDA-based text features as input to a deep neural network to detect automobile insurance fraud. Narayan et al. [2018] introduced a topic-aware convolutional neural network to generate summaries from online news articles. LDA was used to generate document-topic distributions and word-topic distributions separately, and a CNN was then incorporated to encode and decode the document representations.

However, such approaches generate document representations without considering the characteristics of document structure hierarchically. To address this issue, a Hierarchical Attention Network (HAN) [Yang et al., 2016a] has been previously developed, which can capture the hierarchical features on both word level and sentence level through a stacked RNN architecture. In this chapter, the LDA distributions are incorporated into the architecture ofthe HAN model to enhance the word level representation and sentence level representation simultaneously.

6.3 Method

Hierarchical frameworks utilise the document structural features such as the relation between word and sentence, and between sentence and document. Meanwhile, the LDA model generates different distributions which can be used as additional information for encoding document representation. In order to investigate the effectiveness of a learning model which encodes documents hierarchically and incorporates LDA distributions, this section compares different neural network structures with/without the inclusion of LDA distributions. It first establishes three different neural network structures (i.e., CNN, RNN and Transformer) without considering structural features, and then compares these three networks with/without LDA distributions. It also applies two hierarchical models to evaluate the combination of structural features and LDA distributions.

6.3.1 LDA Distributions

The LDA model generates topic-word distribution and document-topic distribution simultaneously. The former is shared between all documents and contains global word co-occurrence features in the whole corpus, while the latter is the local distribution over the topics for a given document, and is independent of all other documents. These two distributions can be used as additional features in the word level and sentence level encoder layer in the hierarchical frameworks. Each sentence is represented by implementing a specific neural network architecture to encode the combination of word embeddings and transposed topic-word distributions. Similarly, the document is then represented by encoding all sentence representations which are generated from the previous step. Finally, the document representation is concatenated with document-topic distribution as an additional feature to make the final prediction.

6.3.2 Model Specifications

Let D denote a document consisting of a sequence of sentences (s_1, s_2 , ... , s_m); Meanwhile, let s_i denote a sentence consisting of words ($w_{s_i}^1$, $w_{s_i}^2$,..., $w_{s_i}^n$) where $i \in [1, m]$, the model embeds s_i into a distributional space $x =$ (x_1, x_2, ... , x_n) where $x_j \in \mathbb{R}^k$, $j \in [1, n]$ and k is the dimension of word embedding. Meanwhile, the LDA model generates topic-word distribution, which are transposed as $tw = (tw_1, tw_2, ..., tw_n)$ where $tw_j \in \mathbb{R}^t$ (t denotes number of topics) and the document-topic distribution can be denoted as $dt = (dt_1, dt_2, ..., dt_d)$ where $dt \in \mathbb{R}^{d \times t}$. All the models are trained to minimize their cross-entropy error:

$$\ell(\tilde{y}) = \sum_{p=1}^{c} y_p \log(\tilde{y}_p) \tag{6.1}$$

where $y, \tilde{y}$ are the ground-truth label and predicted label respectively, c denotes number of classes.

LDA based Non-Hierarchical Models

Three different network structures are implemented as the encoding layers in the LDA-based non-hierarchical models. Figure 6.1 depicts the overall model structure. Formally, each document representation is generated from the initial tokens in the document. This is an aggregation of all the word embeddings x to the encoding layer. Meanwhile, the LDA model also takes text input to

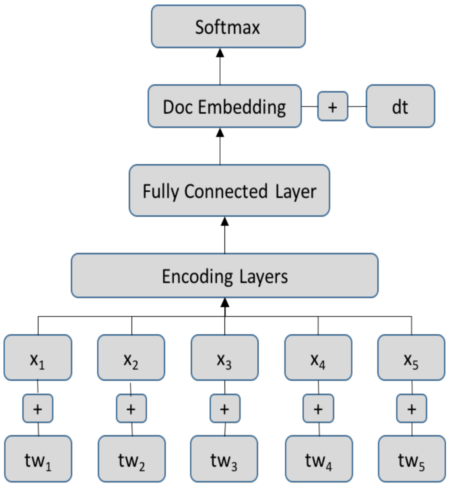


Figure 6.1: LDA based non-hierarchical models structure

generate topic-word distribution and document-topic distribution simultaneously. Next, the transposed topic-word distribution tw is concatenated with word embeddings as the input to the encoding layer. The document-topic distribution dt is then concatenated with the generated document representation. Finally, a Fully Connected (FC) layer with softmax activation and Adam optimizer is made for the final classification.

CNN: For a possible variant CNN structure, Kim's implementation Kim [2014] is adopted as the baseline CNN model. It consists of 128 filters and 3 different convolutional filter sizes $h \in [2,3,4]$ with ReLU activation, with each convolutional layer followed by a max-pooling layer. The results from the max-pooling layers are concatenated, going through a Fully Connected (FC) layer with 50 hidden units. Formally, the convolutional layer using different filter operators $W_{h,j} \in \mathbb{R}^{h \times k}$ is applied to a window of h words to produce a new feature c_j^h at the word level:

$$c_j^h = ReLU\left(\left(x_{j:j+h-1} \oplus tw_{j:j+h-1}\right) \circ W_{h,j} + b_{h,j}\right) \tag{6.2}$$

where the notation $\circ$ and $\oplus$ denote element-wise multiplication and concatenation respectively, $ReLU$ denotes the nonlinear function, $b_{h,j}$ is a bias term. Then, the max-over-time pooling function is used to capture the most important feature $\tilde{c}_j^h$:

$$\tilde{c}_j^h = Max\left(c_j^h\right) \tag{6.3}$$

The final feature maps are formed by concatenating all $c_j = (\tilde{c}_j^1, \tilde{c}_j^2, ..., \tilde{c}_j^h)$, then the document representation d can be generated by a FC layer:

$$d = ReLU\left(c_j \circ W_j + b_j\right) \tag{6.4}$$

where W_j is a weight matrix and b_j is a bias term. Finally, the document representation d is concatenated with dt to make the final prediction in a softmax layer. **Self-Attentive RNN**: This study applies self-Attentive LSTM Lin et al. [2017] as the baseline RNN model. It consists of two LSTMs with 50 hidden units and a dropout of probability 0.2 in each direction. In addition, the self-attention layer has 100 hidden units for the outputs from LSTM, and is then followed by an FC layer with 32 hidden units and ReLU non-linearity.

Formally, the forward $\overrightarrow{r_n}$ and backward $\overleftarrow{r_n}$ hidden states at the word level can be obtained by using bidirectional LSTM:

$$\overrightarrow{r_n} = \overrightarrow{LSTM}\left(x_{1:n} \oplus tw_{1:n}\right) \tag{6.5}$$

$$\overleftarrow{r_n} = \overleftarrow{LSTM}\left(x_{1:n} \oplus tw_{1:n}\right) \tag{6.6}$$

Then the $\overrightarrow{r_n}$ and $\overleftarrow{r_n}$ can be concatenated as $r_n = (\overrightarrow{r_n}; \overleftarrow{r_n})$, thus each document is encoded as $\tilde{r}_n = (r_1, r_2, ..., r_n)$ where $\tilde{r}_n \in \mathbb{R}^{n \times 2u}$ (u is the hidden unit for each unidirectional LSTM), which is then passed to attention mechanism to get annotation matrix α_n:

$$\alpha_n = softmax\left(W_{s2} Tanh(W_{s1} \tilde{r}_n^T)\right) \tag{6.7}$$

where $W_{s1} \in \mathbb{R}^{p \times 2u}, W_{s2} \in \mathbb{R}^{l \times p}$ (p is the number of neuron units, l denotes to use l times attention) are parameters to learn the important components of the document. The annotation matrix $\alpha_n \in \mathbb{R}^{l \times n}$ multiply $\tilde{r}_n$ to compute the l weighted sums to get the final document representation d.

$$d = \sum_n \alpha_n \tilde{r}_n \tag{6.8}$$

Finally, the document representation d is concatenated with dt to make the prediction in a softmax layer.

Transformer: The encoder part of Transformer Vaswani et al. [2017a] is implemented to evaluate its performance on the document classification task. It first calculates the Positional Embeddings (PE) with 300 dimensions for the input, and sum the PE with the original word embeddings instead of concatenation. For the multi-head self-attention, it uses a total of eight heads, where each head has 16 units. It then takes the average of each step of the output sequence from the self-attention layer, followed by an FC layer with 32 hidden units and ReLU non-linearity. Formally, it uses the scaled-dot-product attention to compute the most pertinent information to that document:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{k}})V \tag{6.9}$$

where Q, K, V are 'query', 'key' and 'value' embeddings which concatenate word embeddings x_n with word-topic distribution wt_n. Thus, the final document representation can be formed by multihead attention:

$$\begin{aligned} Multihead(Q, K, V) &= [head_1, head_2, ..., head_n] \\ where \quad head_n &= Attention(Q_j, K_j, V_j) \end{aligned} \tag{6.10}$$

The final output is the concatenation of the outputs from each head, which is then concatenated with dt to make the final prediction in a softmax layer.

LDA based Hierarchical Models

In this section, it utilises two different hierarchical models to investigate the document representation with/without the LDA features. Figure 6.2 depicts the overall hierarchical framework structure. The hierarchical models take word and sentence representation as inputs at different phases. The word-topic distribution tw is concatenated with word embeddings x, and is aggregated to a sentence representation to the encoding layer. The document representation can then be formed by aggregating all the sentence representations s. The document-topic distribution dt is concatenated with the generated document representation. An FC layer with softmax activation and Adam optimizer is used for the final classification.

ESRC: This study implemented a similar structure to the ELMo Sentence Representation Convolutional Network (ESRC) Jiang et al. [2019b] for the hierarchical Convolutional framework, but using the pre-trained GloVe embeddings instead of the ELMo embeddings in order to compare them with other hierarchical models. Formally, the word encoder has 128 filters and 7 different convolutional filter sizes $h \in$ [1,2,3,4,5,6,7] with ReLU activation, followed by a batch normalization and a max-pooling layer. The results from the max-pooling layers are concatenated and passed to an FC layer with 32 hidden units and ReLU activation to form a sentence representation. The sentence encoder takes each sentence representation as the input, with the same struc-

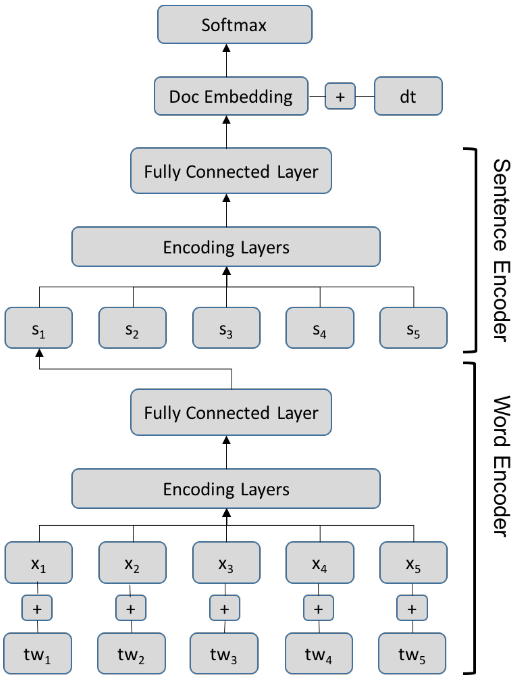


Figure 6.2: Hierarchical model structure

ture as the word encoder. Similar to Kim's CNN, the encoding convolutional layer using different filter operators $W_{h,j} \in \mathbb{R}^{h \times k}$ is applied to a window of h words to produce a new feature $c^h_{x_j}$ at the word level:

$$c^h_{x_j} = BN\left(ReLU\left((x_{j:j+h-1} \oplus tw_{j:j+h-1}) \circ W_{h,j} + b_{h,j}\right)\right) \tag{6.11}$$

where the notation $\circ$ and $\oplus$ denote the element-wise multiplication and the concatenation respectively, $ReLU$ denotes the nonlinear function, $b_{h,j}$ is a bias term; this study also adds a batch normalization BN on top of the convolutional layer.

Then, the max-over-time pooling function is used to capture the most important feature $\tilde{c}^h_{x_j}$:

$$\tilde{c}^h_{x_j} = Max\left(c^h_{x_j}\right) \tag{6.12}$$

The final word-level feature maps are formed by concatenating all $c_{x_j} = (\tilde{c}^1_{x_j}, \tilde{c}^2_{x_j}, ..., \tilde{c}^h_{x_j})$, then the sentence representation s_i can be generated by an FC layer:

$$s_i = ReLU\left(c_{x_j} \circ W_j + b_j\right) \tag{6.13}$$

where W_j is a weight matrix and b_j is a bias term. Then, the final document representation d can be obtained similarly: it first obtain sentence-level feature maps $c^h_{s_i}$ by convoluting the sentence sequence using different filter operators, followed by batch normalization:

$$c_{s_i}^h = \left(c_1^h, c_2^h, ..., c_{s_{i:i+h-1}}^h\right) \tag{6.14}$$

Then, the max pooled features can be obtained:

$$\tilde{c}_{s_i}^h = Max\left(c_{s_i}^h\right) \tag{6.15}$$

Finally, after concatenating all $\tilde{c}_{s_i}^h$ to obtain c_{s_i} the document representation d can be formed as:

$$d = ReLU\left(c_{s_i} \circ W_i + b_i\right) \tag{6.16}$$

where W_i is a weight matrix and b_i is a bias term, $ReLU$ is the non-linear function. Finally, the document representation d is concatenated with dt_i to make final predictions in a softmax layer.

HAN: The Hierarchical Attention Network [Yang et al., 2016a] was implemented for the hierarchical RNN framework. The word-encoder Bi-LSTM has 100 dimensional hidden units with a dropout of probability 0.2. The sentence encoder has the same structure as the word encoder, except that it has an extra FC layer with 32 hidden units and ReLU non-linearity.

Formally, the forward $\overrightarrow{r_{x_n}}$ and backward $\overleftarrow{r_{x_n}}$ hidden states at the word level can be obtained by using bi-directional LSTM:

$$\overrightarrow{r_{x_n}} = \overrightarrow{LSTM}\left(x_{1:n} \oplus tw_{1:n}\right) \tag{6.17}$$

$$\overleftarrow{r_{x_n}} = \overleftarrow{LSTM}\left(x_{1:n} \oplus tw_{1:n}\right) \tag{6.18}$$

Then the $\overrightarrow{r_{x_n}}$ and $\overleftarrow{r_{x_n}}$ can be concatenated as $r_{x_n} = (\overrightarrow{r_{x_n}}; \overleftarrow{r_{x_n}})$. Together with attention matrix α_n, they are used to calculate the importance of each word. The sentence representation s_m is formed by

$$\alpha_n = softmax(W_{n2} tanh(W_{n1} \circ r_{x_n})) \tag{6.19}$$

$$s_m = \sum_n \alpha_n r_{x_n} \tag{6.20}$$

where W_{n1}, W_{n2} denotes the context vector jointly learning the importance of each word in the sentence. Similarly, the document representation d can be also formed by:

$$\overrightarrow{r_{s_m}} = \overrightarrow{LSTM}\left(s_{1:m}\right) \tag{6.21}$$

$$\overleftarrow{r_{s_m}} = \overleftarrow{LSTM}\left(s_{1:m}\right) \tag{6.22}$$

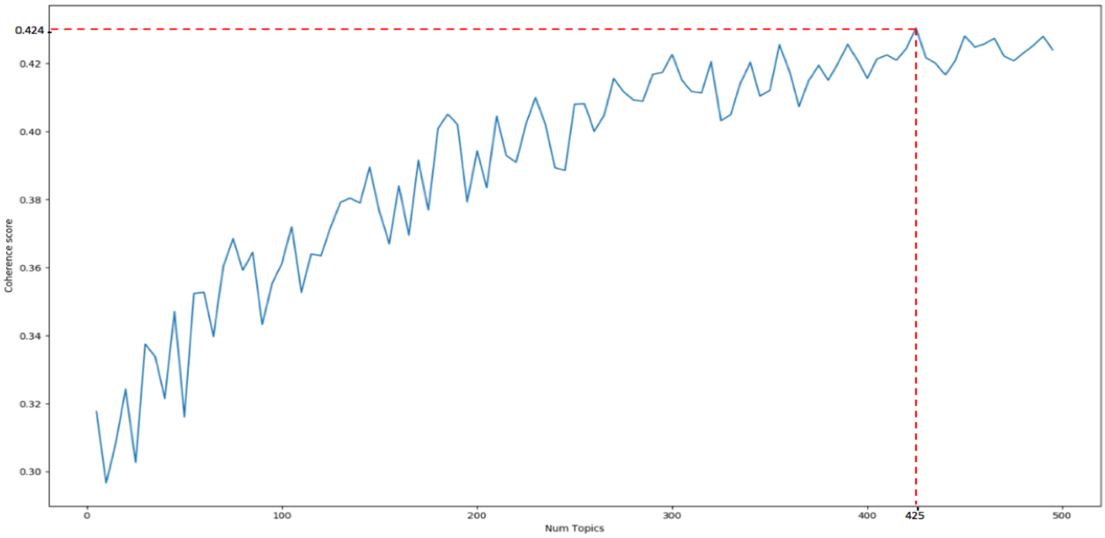


Figure 6.3: Coherence scores in 500 topics

Then the $\overrightarrow{r_{s_m}}$ and $\overleftarrow{r_{s_m}}$ can be concatenated as $r_{s_m} = (\overrightarrow{r_{s_m}}; \overleftarrow{r_{s_m}})$. Together with attention matrix α_m, they are used to calculate the importance of each sentence. The document representation d is formed by

$$\alpha_m = softmax(W_{m2}tanh(W_{m1} \circ r_{s_m})) \tag{6.23}$$

$$d = \sum_m \alpha_m r_{s_m} \tag{6.24}$$

where W_{m1}, W_{m2} denotes the context vector jointly learning the importance of each sentence in the document. The document representation d is concatenated with dt to make final predictions in a softmax layer.

6.4 Experiment

This study splits the dataset into training and test sets with a ratio of 9:1. It performs 10-fold cross-validation on the training set, then fine-tune and obtain the best performing model based on the test set. The final scores are obtained based on the average of 10 predictions on the test set.

6.4.1 Dataset

The Hyperpartisan News Detection dataset[2] contains two parts. The *By-Publisher* corpus contains 750K articles which were automatically classified,

[2] *https://zenodo.org/record/1489920.XcVDj9Hgrew*

based on a categorisation of the political bias of the news provider. The *By-Articles* corpus contains 1,273 articles which were annotated manually. Although the *By-Publisher* corpus has great potential in training deep learning models due to its significant size, Chapter 4 revealed that there is no significant correlation between the two corpora, in the sense that training a learning model on the *By-Publisher* corpus leads to low performance in the task of predicting partisanship on the *By-Article* corpus. Thus, in the evaluation described in this chapter, all models are only trained on the *By-Article* corpus, as this is more reliable based on its manual annotation assessment Vincent and Mestre [2018a], and it is also the official ranking corpus for the task. This study only uses the training set (645 articles) of the *By-Article* corpus, as the rest (628 articles) of the corpus is unavailable to the public (only used for system evaluation).

As discussed previously, the large differentiation in document size makes it impractical to directly use word-level representations as the input, as most news articles have no limitation on sequence length compared to other types of sources (e.g., reviews, tweets, etc). This study uses the same configuration as in the previous chapter, selecting the initial 512 tokens to represent each article in the LDA-based non-hierarchical models. For the hierarchical models, it takes a maximum of 100 words per sentence, and 30 sentences per document.

6.4.2 Pre-processing

As before, this study extracts the title and article text from the original XML file, and represents each article as a sequence of sentences. The text paragraphs are split into sentences, and white spaces are normalised. Since the latest contextualised word embeddings have already been evaluated on this dataset, this evaluation therefore use the traditional pre-trained GloVe model[3] to generate word embeddings for evaluating the effectiveness of LDA distributions, and the Gensim LDA model with 425 topics to generate topic-word distribution and document-topic distribution. It evaluates the topic coherence [Newman et al., 2010] to find the optimal number of topics for our LDA model, as shown in Figure 6.3.

[3] 6 billion words, 300 dimensions

Model	Accuracy (std)
Transformer	.7212 (.0241)
LDA-Transformer	.7156 (.041)
CNN	.7295 (.0374)
LDA-CNN	.7347 (.0499)
Attentive-RNN	.7363 (.0255)
LDA-Attentive-RNN	.7375 (.0289)
ESRC	.7181 (.1509)
LDA-ESRC	.7369 (.1642)
HAN	.7569 (.0232)
LDA-HAN	**.7652 (.0821)**

Table 6.1: Performance comparison between models. The best model accuracy is marked in **bold**

.

6.4.3 Results and Discussion

The results, presented in Table 6.1, show that, on average, the models incorporating LDA distributions outperform the other models. Specifically, the non-hierarchical models have difficulty handling a wide range of document lengths, especially if document sequences are truncated which could potentially cause information loss. Accordingly, the use of hierarchical frameworks, which identify the importance both on the word level and sentence level features by the corresponding encoders, leads to an improvement in accuracy.

Interestingly, the accuracy of the transformer alone is higher than that of the transformer incorporating LDA distributions, although the transformer models are generally lower than others in accuracy. This is potentially because the transformer was designed for the task of machine translation, and therefore this study only uses the encoder part of it with default configurations. This also aligns with other people's findings[4] that the transformer is not the optimal method for document classification tasks. On the other hand, Attentive-RNN achieves the highest accuracy out of all the non-hierarchical models, especially when it incorporates LDA features. However, the ESRC model gets lower accuracy than most of the non-hierarchical models. The accuracy of this is, however, increased by adding LDA features, and the LDA-ESRC models are better than all the non-hierarchical models. This indicates that the hierarchical frameworks incorporating LDA distributions could improve model performance in terms of accuracy. This is also proved by the

[4] `https://github.com/brightmart/text_classification`

Feature	Accuracy (std)
TW+HAN	.7591 (.0474)
DT+HAN	.7637 (.0211)
TW+DT+HAN	**.7652 (.0821)**

Table 6.2: Performance comparison between combinations of topic distributions. TW denotes the topic-word distribution, DT denotes document-topic distribution. The best model accuracy is marked in **bold**

.

LDA-HAN model, which has better accuracy than the HAN model.

Although most of the models can be improved by adding LDA features, the hierarchical frameworks can achieve greater improvement from them. The non-hierarchical models can achieve an improvement of around 0.32% on accuracy, while the hierarchical models can achieve around 1.36% improvement. Specifically, the hierarchical models consider both word-level and sentence-level information separately, and the topic-word distribution enriches the word-level features by adding word occurrence topic distribution through the vocabulary. On the other hand, the document-topic distribution provides local topic distribution, which is independent of all other documents, to increase feature spaces for the final softmax prediction layer, and leads to better accuracy on the document classification task.

In order to investigate the importance of each distribution that improve the model accuracy, this study also analyses each independent feature separately as shown in Table 6.2. The model accuracy is slightly improved by combining the topic-word distribution with word-level encoder. Then, the model takes the document-topic distribution on the sentence-level encoder and achieves more significant improvement. This is expected since the document-topic distribution contains the independent feature distribution between each document and therefore makes the feature space more distinguishable. The word-topic distribution contains global word co-occurrence information, which is less distinguishable.

6.5 Conclusion

This chapter has explored the performance of different popular neural network structures with/without incorporating LDA distributions on the Hyperpartisan News Detection dataset. It has investigated how the hierarchical models take advantage of the structural features of the document to generate a better document representation compared with non-hierarchical models. It has found that the models that include LDA distributions could enrich the feature space by adding global word co-occurrence topic distribution and local document topic probability on word and sentence level respectively.

This study first evaluated the non-hierarchical model with/without LDA features. The results demonstrate that most of the non-hierarchical models improved their accuracy when combined with LDA features, except for the Transformer model. On the other hand, most of the hierarchical models achieved better accuracy than non-hierarchical models, and also showed greater improvement when combined with the LDA. This indicates that the hierarchical model has the advantage of handling longer document sequences and reducing information loss by incorporating structural features in the document. Moreover, the benefits resulting from the LDA distributions can be strengthened in the hierarchical models. Thus, the combination of hierarchical frameworks and LDA distributions could significantly improve model performance in document classification. To sum up, the experimental results from Chapter 4, Chapter 5 and this chapter indicate that the combination of hierarchical framework, contextualized word embedding and additive LDA distributions outperform other baseline models in the hyperpartisan detection task. Consequently, this combination is implemented as the pre-training model for the transfer learning in the next chapter.

Chapter 7

Transfer Learning: From Hyperpartisan News to Tendentious News

In order to detect tendentious framing in the climate change news articles, this chapter makes use of two corpora: the existing hyperpartisan general domain one, and a relatively smaller tendentious one in the domain of climate change built in this study. The idea is to adapt the model from the hyperpartisan one to the smaller domain-specific one, in order to accelerate the model accuracy and also reduce the annotation cost. This chapter first describes the data collection process, and the details of creating the tendentious climate change news corpus. Then, it presents the transfer learning method in detail. It discusses the model selection, pre-training strategies and fine-tuning methods. Finally, it compares the results from before and after the application of transfer learning.

7.1 Corpus Construction

This chapter uses the same dataset of climate change news articles which was described in Chapter 3. Since annotating such a large number of news articles is time-consuming and costly, in this preliminary experiment, 500 articles are randomly selected for the annotation task, with 125 articles belonging to each publisher.

Table 7.1 shows the key statistics of our collected set per publisher (i.e. the total number of articles, the average number of sentences per article, the average number of words per article, and the standard deviation of this). It is

Publisher	no. of articles	av. sentences	av. words (std.)
Guardian	125	31.2	798.3 (479.3)
Independent	125	29.1	654.1 (358.1)
Telegraph	125	34.7	553.6 (494.8)
Times	125	27.6	669.4 (420.7)

Table 7.1: Key statistics per publisher before removing very large/small word counts articles.

interesting that over all publishers, the average is about 30 sentences long, with word counts ranging from 550 words on average at the *Telegraph* to around 800 at the *Guardian*. Also, while the Telegraph has the highest sentence average, it also has the lowest word average. However, the variation of the word counts is large for all publishers, for instance, the maximum/minimum word counts per article in the *Guardian* are 5901/57 respectively. Since some of the articles are too long/short, taking account of the quality and fairness of the annotation, the articles which have very large or very small word counts were removed to avoid assigning an unbalanced task size to the annotators. Finally, a total of 420 articles was selected, where the average number of sentences and words is more balanced, as shown in Table 7.2. The study also demonstrates the improvements of standard deviation of word counts before and after removing those documents. Unfortunately, the corpus cannot be made available for public use due to the restrictions of the LexisNexis policy.

Publisher	no. of articles	av. sentences	av. words (std.)
Guardian	109	30.7	683.3 (386.4)
Independent	113	29.9	641.6 (311.2)
Telegraph	82	31.3	603.6 (407.9)
Times	116	29.7	624.7 (350.6)

Table 7.2: Key statistics per publisher after removing articles with very large/small word counts.

7.1.1 Gold Dataset

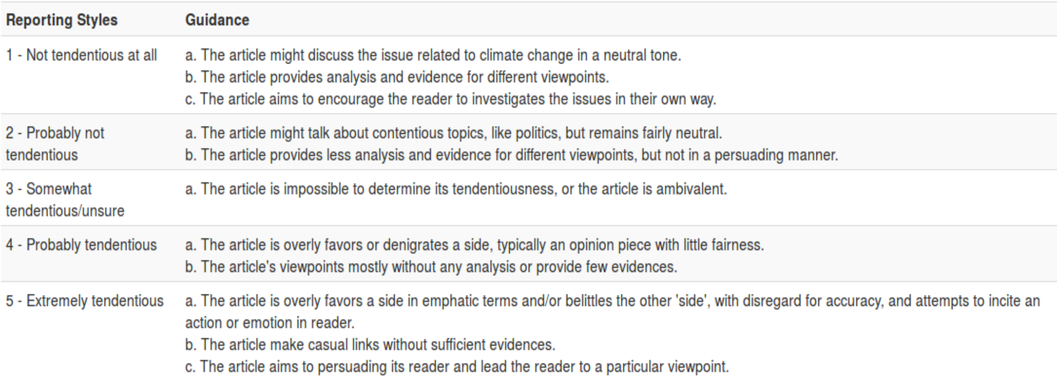
Initially, this study drew articles from a pilot study, representing a corpus of 40 articles as the gold standard dataset. The purpose of this is to pre-assess the reliability of contributing annotators: only those annotators who passed a test on the gold dataset (i.e. they got the same result as the gold standard) would be eligible to proceed to the formal annotation. 10 articles were randomly selected from each publisher and two experts (i.e., researchers familiar with the topic and the concept of tendentiousness) were asked to assign a category to each article. Then for each of these articles, the values provided by the two experts were assimilated. If the two values were judged to be similar, they were merged (following the procedure described in the next section). If the two values were conflicting, the results were discussed to get an overall agreement between annotators.

7.1.2 Annotation Scheme

The Amazon Mechanical Turk[1] was used to present the articles to annotators, who were asked to read each article's web page. Good practice in annotation crowdsourcing encourages presentation of a task in such a way that annotators only have to make easy decisions (e.g. binary decisions), and it has been shown that simpler design without too much variance tends to lead to better results Sabou et al. [2014]. For instance, many sentiment annotation tasks on Twitter, which annotate each tweet as either positive or negative Saif et al. [2012] or which add neutral as a third option Kouloumpis et al. [2011], Wang et al. [2012], have used a simple scheme to keep the task intuitive, as sentiments are normally fairly explicit in tweets. However, a simple annotation scheme might also make it difficult for the annotators to reach a decision and/or to be confident about their decision, especially when the characteristics that define the target are too subtle to be identified from the source. In the pilot experiment of this task, a simple scheme was provided to annotators to identify whether the article was 'True', 'False' or 'Neutral' with respect to being tendentious. However, 66.7% of results from the annotators were assigned as 'Neutral', with 13.2% and 20.1% being assigned as 'True' and 'False' respectively. In other words, the annotators largely found it too difficult to decide whether an article was tendentious or not. Feedback from the annotators also indicated that they were often not confident enough to choose True or False, and therefore chose the safe option of Neutral.

In order to tackle this issue, this study uses a similar annotation scheme to that of Vincent and Mestre [2018b], consisting of a five-point scale to allow annotators to express their degree of certainty, as shown in Figure 7.1. On this scale, the medium value 3 is reserved for when they are unsure about the tendentiousness of article, while the values 1 and 5 represent high confidence that the article is non-tendentious or tendentious respectively, and the values 2 and 4 represent lower confidence that the article is non-tendentious or tendentious

[1] *https://www.mturk.com/*



Reporting Styles	Guidance
1 - Not tendentious at all	a. The article might discuss the issue related to climate change in a neutral tone. b. The article provides analysis and evidence for different viewpoints. c. The article aims to encourage the reader to investigates the issues in their own way.
2 - Probably not tendentious	a. The article might talk about contentious topics, like politics, but remains fairly neutral. b. The article provides less analysis and evidence for different viewpoints, but not in a persuading manner.
3 - Somewhat tendentious/unsure	a. The article is impossible to determine its tendentiousness, or the article is ambivalent.
4 - Probably tendentious	a. The article is overly favors or denigrates a side, typically an opinion piece with little fairness. b. The article's viewpoints mostly without any analysis or provide few evidences.
5 - Extremely tendentious	a. The article is overly favors a side in emphatic terms and/or belittles the other 'side', with disregard for accuracy, and attempts to incite an action or emotion in reader. b. The article make casual links without sufficient evidences. c. The article aims to persuading its reader and lead the reader to a particular viewpoint.

Figure 7.1: Annotation instructions

respectively.

It should be noted that the correlation between confidence in the judgement and strength of tendentiousness is not clear-cut. On the one hand, a moderately tendentious article is theoretically not the same as having only a moderate degree of confidence that the article is tendentious. However, in practice, these two things tend not only to be quite similar, but actually do not matter since the scores for 4 and 5 will be merged into a single "tendentious" score at the end of the process (as will be explained). So the annotators are free to think of a score of 4 either as being moderately tendentious, or in terms of their own moderate degree of confidence. This was made clear to the annotators during their period of instruction. In hindsight, however, it notes that this may still have caused some confusion to the annotators, even if it did not affect the ultimate judgements after merging.

The process of conflict resolution was established as follows. Each article is assigned to two different annotators individually, and their scores are aggregated if they both agree or disagree about the tendentiousness of the article. Specifically, the article is labelled as tendentious if both annotators give 4 or 5, or one of each. Similarly, the article is labelled as non-tendentious if both annotators give 1 or 2, or one of each. If two annotators disagree more substantially (e.g., one gives 4 and another gives 1), this article is kept until the end of task and reassign it to new annotators. If the scores still disagree in the second assignment, it will be reassigned one more time. The articles are rejected from the corpus if disagreement still exists in the reassignment, and add new articles from the larger unused set to make up the numbers.

In total, the task was completed using four annotators, split into two pairs. The corpus was split into two equal sets of 250 articles each, randomly drawn from different UK broadsheets. Each pair of annotators was then assigned one half of the corpus, so that the entire corpus was double-annotated. The annotators were not specifically told which paper the article came from. The pilot annotation ran from 16-27 Sept 2019, including the construction of the gold data set and the test annotation for the other annotators. The formal annotation task ran from 29 September - 20 November 2019.

Following the merging process, the final agreement rate between two annotators is 68.37% on average. 39.6% of the articles were assigned to Non-Tendentious, 29.4% were assigned to Tendentious, and 31% were assigned to the "unsure or somewhat tendentious" category. It can be seen that compared with the results of the binary scheme, the five-point scale allowed annotators to express their degree of certainty much more easily, despite it being more complicated. For instance, before merging, the ratings of 2 and 4 were assigned to 81% and 74% of the annotations respectively. This indicates that most articles still produce significant uncertainty to annotators so that they might be less confident about selecting rate 1 and rate 5. As noted above, the slightly confusing guidelines may also have contributed to this aspect, even though this should not have affected the final rating.

Imbalanced labelled data is a typical problem for many learning models Bauder et al. [2018], Bauder and Khoshgoftaar [2018]. Specifically, the learning models might over-classify the majority class because of its increased prior probability. Consequently, the instances which belong to the minority group are more likely to be misclassified than those which belong to the majority group. In this annotation task, the final corpus contains more non-tendentious than tendentious articles. As this data imbalance could cause the learning model to overclassify the non-tendentious articles, this study then assigned more news articles from the collected corpus until the amount of both categories became exactly balanced. Thus in the final version as shown in Table 7.3, the two classes have an equal number of articles, with 198 in each class, and 396 articles in total. Since the tendentious detection considers all the tendentious news articles from different publications as a whole, the unequal numbers of articles in each publication does not affect the model performance as long as the total number of tendentious news articles is equal to the number of non-tendentious ones.

	Tendentious	Non-tendentious	number of articles
Independent	67	45	112
Telegraph	42	51	93
Guardian	35	53	88
Times	54	49	103
Total	198	198	396

Table 7.3: Final numbers of news article in each category and publication.

7.2 Transfer Learning

Transfer learning is a machine learning technique which enables a pre-trained learning model for one task to be reused as the starting point for a new model on a second task Pratt [1993]. The balanced tendentious corpus could thus be used as the second task in the transfer learning, to accelerate model accuracy. On the other hand, the balanced tendentious corpus could also be implemented as a standalone training corpus for most traditional machine learning techniques (e.g. Support Vector Machine, Logistic Regression, etc.) given its size and the balanced nature of the data. Consequently, this corpus provides new options for different machine learning-based techniques for the task of discovering tendentiousness in the climate change news.

According to the experimental results previously discussed, several hyperpartisan-sensitive models were trained from the SemEval2019 hyperpartisan news corpus. Different deep learning models are implemented, and 10-fold cross validation is used to evaluate their performance. In terms of possible model structures, this study uses two types of hierarchical models, ELMo Sentence Representation Convolutional Network (ESRC, see Chapter 4) and Hierarchical Attention Network (HAN, see Chapter 5), to build the document representation, since they have the best performance for the detection of hyperpartisan news. This study also implements the pre-trained BERT to generate contextualised word embeddings since the expeirmental result of BERT embeddings to be generally better than ELMo, whilst encoding word and sentence hierarchically (see Chapter 5). The LDA distributions are also utilised as additive feature representations for accelerating model accuracy (see Chapter 6). In order to transfer from the hyperpartisan domain to the tendentious one, the hyperpartisan-sensitive models are then fine-tuned based on our balanced tendentious corpus. Finally, a 10-fold cross validation is implemented to evaluate the performance of detecting tendentious framing in the news articles.

7.2.1 Method

In order to compare the effectiveness of transfer learning, the ESRC and HAN are initially trained on the tendentious corpus only to get *ESRC-base* and *HAN-base* as baseline models.

Since hyperpartisan news is similar to tendentious news (see Chapter 2), it is therefore assumed here that their data distributions are also similar. Then, two pre-training strategies are used for investigating data distributions of two corpus:

1. **In-domain pre-training:** Only the hyperpartisan corpus is used as pre-training data, and then the pre-trained hyperpartisan-sensitive model is fine-tuned separately on the tendentious corpus. In this approach, the hyperpartisan set and the tendentious set are trained separately.

2. **Cross-domain pre-training:** The hyperpartisan corpus is integrated with the tendentious corpus to enlarge the training data size as a whole. and a generalised model is trained based on both hyperpartisan news and tendentious news. In this approach, the hyperpartisan set and the tendentious set are trained together.

For fine-tuning the pre-trained models, either the pre-trained model's last layer can be truncated or the weights of the first few layers can be frozen [Yosinski et al., 2014, Howard and Ruder, 2018]. The last layers from deep classifiers are typically used for predicting the labels, so the last layer can be replaced based on the needs of different tasks. For instance, if one task has five labels but the target task has only two, a new layer will therefore replace the original layer which has five labels. Since the hyperpartisan corpus and tendentious corpus have the same number of labels (i.e., True or False), this study implements the fine-tuning by freezing the weights from the top layers of the pre-trained models, since the first few layers capture universal features.

This study evaluates both pre-training strategies by implementing 10-fold cross validation. For in-domain pre-training, it uses the hyperpartisan corpus for pre-training the model, and splits the tendentious corpus into a training set and test set with a ratio of 9:1 for evaluating the accuracy. For cross-domain pre-training, it first integrates the two corpora and then directly split the mixed corpus into a training set and test set with a ratio of 9:1 for cross validation. All results are calculated by averaging the accuracy of test sets of 10-folds.

7.2.2 Result and Discussion

Models	Accuracy (std)
ESRC-base	.7342 (.0961)
HAN-base	.7401 (.0369)
In-ESRC	.7623 (.1042)
In-HAN	**.7759 (.0572)**
Cross-ESRC	.7563 (.1197)
Cross-HAN	.7530 (.0515)

Table 7.4: System comparison, 'base' denotes the baseline models, 'In' denotes the In-domain pre-training, 'Cross' denotes the Cross-domain pre-training.

The results, presented in Table 7.4, show that, on average, the model with transfer learning outperforms the baselines. This indicates that transfer learning could improve the model performance when the target sources suffer from data insufficiency issues.

Such an improvement also demonstrates that the feature distribution in the hyperpartisan corpus is relatively similar to that of the tendentious corpus. However, the similarity is not as high as expected. Since the cross-domain pre-training strategy achieves lower accuracy than in-domain pre-training models, this indicates that the feature distributions from the two corpora are slightly different. The in-domain pre-training model is first trained on hyperpartisan news so that the model weights are updated by capturing hyperpartisan-sensitive distributions. Then, the hyperpartisan-sensitive model is fine-tuned by freezing the weights of its top layers. This implementation enables the model to update no weights on the top layers, but only to update the weights of the last predicting layer during the training of tendentious news. On the other hand, the cross-domain pre-training is trained on both hyperpartisan news and tendentious news. Although the training data is enlarged by this integration, the different feature distributions potentially increase the difficulty of modelling both types of news. This is probably because the hyperpartisan news is generally related to political domains, but the tendentious one is more specifically oriented to climate change. Also the tendentious one has less obvious partisanship, although alarmism and sarcasm can be still found (e.g. "Throughout history, there have been false alarms: "shadow of the bomb", "nuclear winter", "ice age cometh" and so on. So it's no surprise that today many people are skeptical about climate change. The difference is that we have hard evidence that increasing temperatures will lead to a significant risk of dangerous repercussions."[2]).

7.3 Case Study: Sentiment Analysis in the Tendentious Corpus

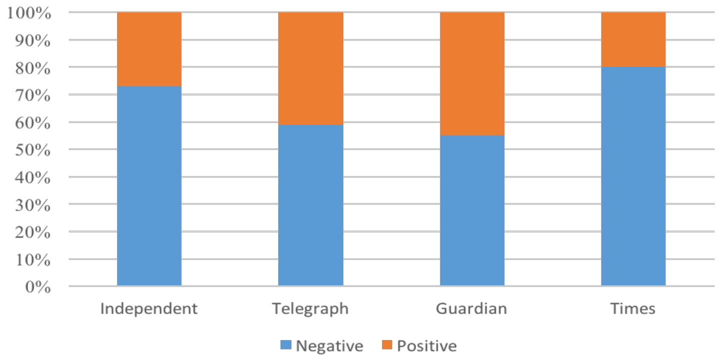
This section establishes a small case study analysing the sentiment orientation in the tendentious news and non-tendentious news. Since tendentious framing typically takes advantage of emotional elements to exaggerate or understate human interest stories [Harrison, 2008], it is assumed that the sentiment orientation of news articles might be affected in such a way as to overly express either a positive or negative tone through such framing. To verify the assumption, it use a similar sentiment analysis method to the one conducted in Chapter 3, since the tendentious corpus does not have any sentiment labels (i.e., positive, negative or neutral).

Specifically, SentiWordNet is implemented, which is the same tool used in Chapter 3, to automatically assign sentiment labels for each article in the tendentious corpus. In order to compare how sentiments vary between those four publications (i.e., *The Guardian, The Telegraph, The Independent* and *The Times*) in tendentious framing and non-tendentious framing, the ratio of positive and negative news articles is calculated for each news publication among each category since each publication has a disproportional number of articles in each category as shown in Table 7.3.

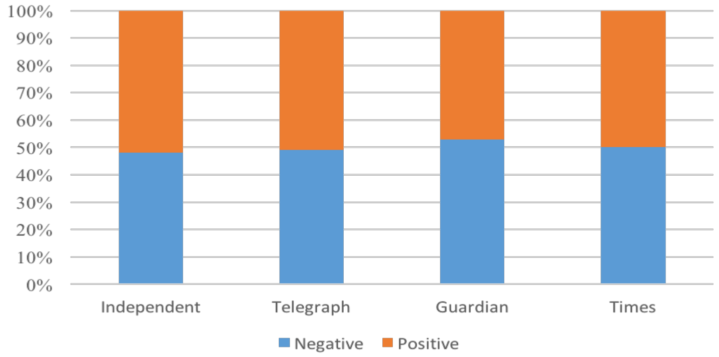
Although SentiWordNet straightforwardly assigns a sentiment label by calculating the frequency of sentiment words (e.g., 'happy', 'sad', etc.) in each news article, this label indicates the overall sentiment orientation for each article generally, as shown in Figure 7.2.

Specifically, the proportion of negative sentiment is higher than that of positive sentiment in the tendentious news articles for all broadsheets, and especially for *The Independent* and *The Times*, as shown in Figure 7.2a. *The Guardian* has the lowest number of negative articles but it is still over 50%. However, non-tendentious news articles convey both positive and negative sentiments, and the overall tone is therefore relatively neutral as shown in Figure 7.2b. This verified our assumption that tendentious framing affects the sentiment orientation of news articles to overly express negative tone. Since the tendentious corpus is related to climate change aspects, such tendency is

[2] Times, 19 April 2009,p.32



(a) Tendentious News Sentiment Ratio



(b) Non-Tendentious News Sentiment Ratio

Figure 7.2: Sentiment Ratio in Tendentious and Non-Tendentious News between Four Publications.

expected as they might have a higher chance to contain negative narrative, such as alarmism. Meanwhile, tendentious framing typically aims to attract and persuade readers in a campaigning and universalistic way, and also negative or loss framing induces more powerful persuasion than positive framing [Meyerowitz and Chaiken, 1987, Lin and Yang, 2014].

To conclude, the sentiment orientations are found to be more disproportionally distributed in the tendentious news, and the non-tendentious news are more balanced in terms of sentiment. Although this is a small case study analysing sentiment orientation between tendentious and non-tendentious news, and the implementation of SentiWordNet is a rather simple method for conducting sentiment analysis, it nevertheless demonstrates some characteristics of tendentious framing in the climate change news articles and also shows the variations of reporting style in different news publications. The tendentious corpus demonstrates the potential for conducting more interesting studies, for instance, combining the sentiment with topic model which was used in Chapter 3 to identify the sentiment orientation towards certain climate change topics.

7.4 Summary

This chapter conducted experiments to investigate the effectiveness of transfer learning for automatically detecting tendentious news about climate change. The experimental findings indicate that the task of annotating tendentious news is more complex than other types of annotation, such as sentiment analysis, since the tendentiousness in the news articles is more implicit. This can be partially explained by the fact that most of the articles produced significant uncertainty to the annotators (the ratings of 2 and 4 were assigned to 81% and 74% of the annotations respectively before merging). Also, the in-domain pre-training strategy was able to significantly enhance the model performance, and transfer learning helped the learning model to tackle the data insufficiency issue when their domains are similar. This chapter also compared the different sentiment orientations between the broadsheets in each category. The results indicate that the tendentious news articles are generally more unbalanced in terms of sentiment since tendentious framing tries to emphasise a certain side of the issue and downplay others. The proportion of negative articles in the different newspapers is also different in tendentious news, where *Times* and *Guardian* have the highest and lowest ratio of negative sentiment respectively.

Chapter 8

Conclusion and Future Work

This thesis has presented some novel approaches and findings on detecting news attitudes and journalistic framing on news reporting about climate change. Due to the controversiality of the climate change issue, which leads to strong polarisation of opinions, this makes it an ideal topic for comparing how different news media vary in their reporting of climate change issues.

To address this, this thesis investigated two factors: news attitudes and journalistic framing. In terms of the former, identifying the attitudes presented in newspaper articles is difficult, as the topics mentioned in the news are quite varied, and documents can be long with many different points of view mentioned throughout each article.

Therefore, this thesis combined a topic model, Latent Dirichlet Allocation (LDA), with opinion mining techniques to automatically identify the topics of the news articles and the opinions towards those topics correspondingly. Experimental results indicated that, as expected, different news publications present different angles towards certain climate change issues.

However, although the combination of LDA with opinion mining enables us to identify the attitudes from different publications, it only tells us whether a piece of news is positive or negative. Thus, the focus was extended to the problem of journalistic framing, where news media in a polarised environment may intentionally frame news to promote a certain side of the issue, and downplay other facts of climate change. However, traditional journalistic framing detection relies on human assessments, and therefore is limited to small scale studies. To address this problem, this work takes advantage of the recent growth of media digitisation, which has created large amounts of online news offering valuable opportunities for implementing machine learning techniques on such data. This enables methods to be developed for automatically identifying journalistic framing, which is a relatively novel topic. However, there remains another challenge with this approach: for machine learning, training data needs to be created, but it is time-consuming and costly to manually annotate news articles on a large scale for this purpose.

In this thesis, to address the above issues, transfer learning techniques were implemented. These take advantage of an available source dataset, train a model on that, and then transfer the knowledge to a relatively smaller target dataset in order to reduce the data insufficiency issue. First the similarity was investigated between two kinds of journalistic framing: hyperpartisan news and tendentious news. Then, the ESRC model to accurately identify the hyperpartisan news automatically was introduced. In order to optimise the model performance, the ESRC model was also extended to a set of its variants by incorporating contextual word embedding with hierarchical frameworks and using distributions generated from an LDA model to enhance the model accuracy. This thesis also established a tendentious climate change news corpus using crowdsourcing, and applied transfer learning based on the model trained on the hyperpartisan news, evaluating different pre-training and fine-tuning strategies. Experimental results indicated that transfer learning could significantly improve the model performance for detecting tendentious news.

8.1 Addressed Challenges

In order to answer the research questions in Chapter 1, this thesis tackled the research challenges in several ways:

1. **Context Complexity**: Since news articles are naturally more flexible in terms of number of paragraphs and variation in document size than other types of text resources (e.g., tweets, micro-blogs, etc.), it is difficult to identify the main viewpoint in a news article where several views about different topics are expressed. The learning model needs to have the ability to adapt to various document sizes whilst encoding document representation. This thesis first implemented the topic model, LDA, to automatically identify topics from news articles, and regroup similar topics that the news articles discuss, based on the topic distributions of those articles. This implementation allows us to understand which news articles have similar or the same topics (which form the opinion targets in this study). Then, opinion mining techniques were applied to analyse the sentiment towards the identified opinion targets. Meanwhile,

hierarchical frameworks were developed which take account of the document's structural information between word and sentence, and between sentence and document. This thesis extensively compared the capacity of a hierarchical framework to encode long documents with traditional encoding methods. The experimental results indicated that hierarchical frameworks could significantly improve model performance in terms of recall and accuracy.

2. **Costly Annotation**: Deep learning methods for detecting journalistic framing in climate change news require an annotated dataset for training the learning models. However, annotating a large volume of news articles is both time-consuming and costly. To avoid over-fitting of the learning models on a small dataset, this thesis established a tendentious climate change news corpus and applied transfer learning, transferring the knowledge from a pre-existing hyperpartisan news corpus to the new tendentious news corpus. It investigated various pre-training strategies and fine-tuning methods, and found that transfer learning with in-domain pre-training could improve model performance on the task of detecting tendentious news. Finally, the transfer learning method also minimises the annotation cost, since the tendentious news corpus is relatively small compared with others (e.g., the hyperpartisan news corpus).

3. **Model Optimization**: In order to optimise the ability of model generalisation, this thesis intensively investigated different encoder structures (i.e., RNN, CNN, DAN, Transformer), different hierarchical frameworks (i.e, ESRC, HAN), and different feature representations (i.e., ELMo, BERT and LDA). Since news articles typically have various document sizes, hierarchical frameworks turned out to outperform other neural architectures, while the recent pre-trained language model BERT generates more effective word representation than other models. LDA distributions were also applied as additive features and it was found that this could also improve model generalisation. Based on the above observations, the model is able to accurately identify the journalistic framing in the news articles.

8.2 Research Questions

For the questions listed in Chapter 1, this section summarise the findings for each one:

RQ1: *How to automatically analyse the overall attitude of newspaper article towards a certain climate change issue?*
This thesis combines the LDA topic model with SentiWordNet to automatically identify the main topic and assign sentiment to that topic correspondingly. Specifically, it selects the topic which has the highest probability among other topics that are generated from LDA, and then implement SentiWordNet to calculate the frequency of sentiment words in each article. The experimental result indicated that different news publications have different angles and attitudes whilst reporting about certain climate change issues. In Chapter 3, the thesis demonstrated how four UK broadsheets reported differently about the Copenhagen Summit in 2009 in terms of topics and sentiment. Generally, *The Guardian* is the most positive and *The Independent* also tends towards the positive, while *The Times* is the most negative and *The Telegraph* is slightly negative. This study also manually investigated some of the articles and found some clear indications that match the automatic analysis results.

RQ2: *How to automatically and accurately detect journalistic framing from climate change news?*
Although opinion mining allows the analysis of the overall positive and negative attitudes of news publications toward certain climate change issues in a positive or negative way, such binary indications can not fully demonstrate how the news article is structured, for instance, how a article is framed to promote a certain side of the issue and downplay others. On the other hand, news article typically have diverse context length, while traditional learning models imply either the maximum sequence length is used to fully represent the longest document, which causes a high computational cost, or alternatively a significant information loss if the sequence length is restricted to a manageable number of initial tokens from the documents. In order to effectively separate biased news from more balanced ones, this thesis developed a method used to participate in SemEval2019 Task 4: Hyperpartisan News Detection in Chapter 4, consisting of the ESRC model which takes account of each article as a set of sequences, and incorporates with ELMo word embeddings to generate effective document representation for news articles. The experimental results demonstrated that the ESRC model could detect the hyperpartisan news effec-

tively and outperformed others in terms of accuracy in the hyperpartisan news detection task. To detect the journalistic framing from climate change news, this thesis also implemented a transfer learning method by taking account of the similarity between hyperpartisan news and tendentious news. Chapter 7 intensively investigated different pre-traininng strategies for transfer learning, as well as fine-tuning methods. The experimental result indicates that transfer learning with in-domain pre-training could significantly improve model accuracy when the tendentious corpus is relatively small. Also, in order to compare how different news media vary in their reporting of climate change issues, a similar sentiment analysis was implemented in Chapter 3. It was found that tendentious climate change news has a more disproportional number of articles in terms of negative sentiment, but the tone of non-tendentious ones is relatively more balanced.

RQ3: *What are the optimal structures and configurations of the learning models in detecting journalistic framing in the news articles?*
The ESRC model demonstrates the potential of combining a hierarchical framework with contextualized word embedding. This study extends this idea to a further step towards achieving the optimal model accuracy. Chapter 5 presented the hierarchical frameworks by taking account of the structural information between words and sentences, and between sentences and document. Meanwhile, the recent pre-trained language models ELMo and BERT are also used to generated contextualized word embeddings. The combination of hierarchical frameworks and contextualized word embeddings outperform other baselines in the hyperpartisan news detection task. Also, this work takes a further step to investigate the model optimization by taking account of the additive feature representation generated from LDA distributions, as explained in Chapter 6. The topic-term distribution and document-topic distribution from the LDA model can be integrated in word level representation and sentence level representation simultaneously. Such implementation also improves the model performance for detecting journalistic framing.

8.3 Limitations

There are also some limitations in this thesis that are listed as follows:

- Although the combination of LDA topic model with SentiWordNet could automatically identify the sentiment towards the topic, the evaluation method is restricted to manual assessment. This is problematic when the study is scaled up. Meanwhile, SentiWordNet assigns a sentiment label for each article by simply calculating the term frequency, this will also potentially miss some useful information, for instance, negations. Also, SentiWordNet is not ideal because it does not cover all the terminology in the specific domain of climate change, nor does it deal with context. A potential solution would be to develop a semi-supervised learning approach, based on a small corpus of manually annotated sentiment news articles, to build a weak classifier to bootstrap the sentiment labels for unlabelled data. Since this study could obtain the weak label for each article, the sentiment analysis could be extended to build supervised learning models which could capture semantic features that could be missed in the SentiWordNet. Consequently, this could automatically evaluate the model accuracy through the weak labels.

- The thesis used the sentiment polarity for detecting the attitude of news article. Although the emotional state of the news article could partially reflect to the viewpoints in the news article, Aldayel and Magdy [2019] have found that there is a disparate alignment between the sentiment and the stance. Consequently, using sentiment polarity as solely feature to detect the attitude might not able to fully capture the stance of the article. In future work, this thesis will also look into the stance in the news article, and merge it with the sentiment feature to enhance the detection.

- This thesis implemented the language model BERT as an embedding generator such as ELMo and Word2Vec; The hierarchical frameworks are then built on the top of the BERT embeddings. Although this provides a potential solution for combining BERT embedding with various neural encoders, this study did not directly compare the result with the performance of the fine-tuning pre-trained models since the fine-tuning method has also demonstrated state-of-the-art performance in document classification [Sun et al., 2019]. Future work will also evaluate the effectiveness of the method with fine-tuned language models such as BERT, GPT-2, etc.

- This thesis conducted a relatively simple sentiment analysis for investigating how news media vary in reporting about climate change without considering syntactic rules and semantic features. Also, the sentiment analysis is purely based on the SentiWordNet lexicon, and this might potentionally cause an out-of-vocabulary problem if a sentiment word is not included in the lexicon. Meanwhile, the scale of the hyperpartisan corpus and tendentious corpus are also relatively small and could potentially cause over-fitting to the learning models. This existing method could therefore be extended to other corpora for evaluating the model generalisation.

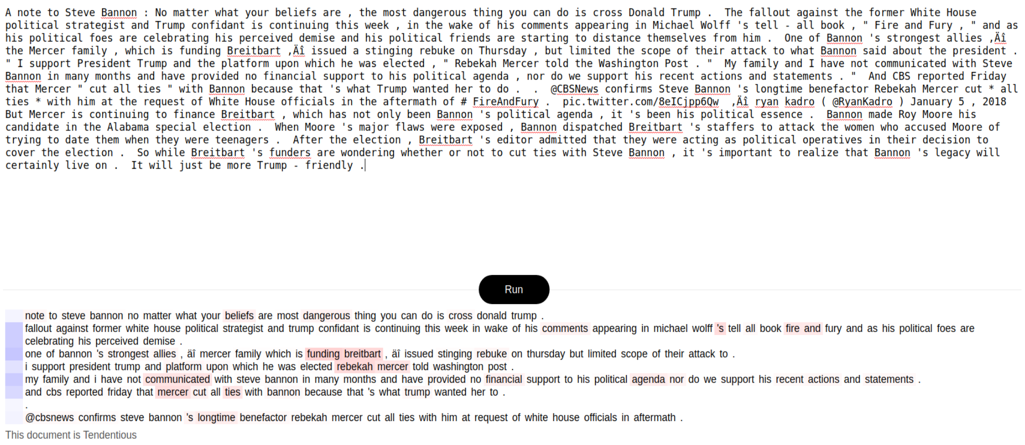
8.4 Future Work

As suggested by the limitations above, the current research work can be extended in several ways in future work:

- **Generalising to other corpora**: While the transfer learning demonstrated its potential in improving model accuracy for detecting journalistic framing, the current transfer learning experiment is limited to two types of frames (i.e., hyperpartisan framing and tendentious framing). The experiment could be extended to other types of frames that might be encountered on any issue of public concern. The Media Frames Corpus (MFC) [Card et al., 2015] offers 15 generic journalistic frames across different aspects of social concerns, and also contains a large volume of data. Therefore, applying transfer learning on the MFC could potentially improve the ability of model generalisation on climate change-related news. Future work could evaluate the models on the MFC and then apply transfer learning for climate change, or other subjects in news frames detection.

- **Adaptation to pre-trained language models**: BERT and its variants (e.g., RoBERTa [Liu et al., 2019b], DistilBERT [Sanh et al., 2019], etc.) have demonstrated their state-of-the-art performance in several NLP tasks. This thesis has focused on using the pre-trained language models to generate embeddings, and training such embeddings based on hierarchical frameworks with different neural encoders, rather than by directly fine-tuning them. In the future work, the pre-trained language models could be evaluated by fine-tuning them based on the journalistic framing corpus.

- **Explainable framing detection models**: One of the limitations of deep learning methods is their black-box nature, making it difficult to understand which aspects of the input data drive the decisions of the



```
A note to Steve Bannon : No matter what your beliefs are , the most dangerous thing you can do is cross Donald Trump .  The fallout against the former White House
political strategist and Trump confidant is continuing this week , in the wake of his comments appearing in Michael Wolff 's tell - all book , " Fire and Fury , " and as
his political foes are celebrating his perceived demise and his political friends are starting to distance themselves from him .  One of Bannon 's strongest allies ,Äî
the Mercer family , which is funding Breitbart ,Äî issued a stinging rebuke on Thursday , but limited the scope of their attack to what Bannon said about the president .
" I support President Trump and the platform upon which he was elected , " Rebekah Mercer told the Washington Post . "  My family and I have not communicated with Steve
Bannon in many months and have provided no financial support to his political agenda , nor do we support his recent actions and statements . "  And CBS reported Friday
that Mercer " cut all ties " with Bannon because that 's what Trump wanted her to do .  .  @CBSNews confirms Steve Bannon 's longtime benefactor Rebekah Mercer cut * all
ties * with him at the request of White House officials in the aftermath of # FireAndFury .  pic.twitter.com/8eICjpp6Qw  ,Äî ryan kadro ( @RyanKadro ) January 5 , 2018
But Mercer is continuing to finance Breitbart , which has not only been Bannon 's political agenda , it 's been his political essence .  Bannon made Roy Moore his
candidate in the Alabama special election .  When Moore 's major flaws were exposed , Bannon dispatched Breitbart 's staffers to attack the women who accused Moore of
trying to date them when they were teenagers .  After the election , Breitbart 's editor admitted that they were acting as political operatives in their decision to
cover the election .  So while Breitbart 's funders are wondering whether or not to cut ties with Steve Bannon , it 's important to realize that Bannon 's legacy will
certainly live on .  It will just be more Trump - friendly .
```

note to steve bannon no matter what your beliefs are most dangerous thing you can do is cross donald trump .
fallout against former white house political strategist and trump confidant is continuing this week in wake of his comments appearing in michael wolff 's tell all book fire and fury and as his political foes are celebrating his perceived demise .
one of bannon 's strongest allies , äî mercer family which is funding breitbart , äî issued stinging rebuke on thursday but limited scope of their attack to .
i support president trump and platform upon which he was elected rebekah mercer told washington post .
my family and i have not communicated with steve bannon in many months and have provided no financial support to his political agenda nor do we support his recent actions and statements .
and cbs reported friday that mercer cut all ties with bannon because that 's what trump wanted her to .
.
@cbsnews confirms steve bannon 's longtime benefactor rebekah mercer cut all ties with him at request of white house officials in aftermath .

This document is Tendentious

Figure 8.1: Example of the importance of words and sentences in a news article. The blue and red color respectively denote the importance of sentence and word, the darker the more important.

network. On the other hand, there is a high level of complexity in framing analysis that often requires a careful investigation of nuances in news coverage. Therefore, framing detection models need to be explainable in order to understand the reasons for certain decisions in the model making. Visual analytics for deep learning models has played an important role in providing an in-depth understanding of how deep learning models work [Choo and Liu, 2018]. In the future work, an attention mechanism will be implemented for capturing the insights of the news article. Beside simply training a deep learning model, an attention mechanism allows deep learning models to focus on the most salient information in the feature representations, and to quantify such importance by involving attention weights. Such attention weights can be visualised to understand, for example, which words or sentences contribute the most attention for the model to make certain decisions.

This thesis made a preliminary demo for this purpose, as shown in Figure 8.1. In this demo, it can be visualised how the decisions of the model are made for the given prediction. For instance, given a piece of text from a newspaper article (i.e., the text box above 'Run' button), the model will first post-process the text by removing stop-words and punctuation, and by lower casing. Then, it will take the word and sentence as inputs separately to the model. In the output box (i.e., the text box below 'Run' button), the original news article has been split into sentences, the darker blue on the right hand side denotes the importance (i.e., the attention weights) of the sentence to the model in order make such a prediction (i.e., shown at the bottom of the figure). Similarly, the darker red highlights in each sentence denote the importance of a word to a model for making such a prediction.

Appendix A

Comparing Attitudes to Climate Change in the Media using Sentiment Analysis based on Latent Dirichlet Allocation

Comparing Attitudes to Climate Change in the Media using sentiment analysis based on Latent Dirichlet Allocation

Ye Jiang[1], **Xingyi Song**[1], **Jackie Harrison**[2], **Shaun Quegan**[3], **and Diana Maynard**[1]

[1]Department of Computer Science
[2] Department of Journalism Studies
[3]School of Mathematics and Statistics
University of Sheffield, Western Bank, Sheffield, S10 2TN, UK
{*yjiang18,x.song,j.harrison,s. uegan,d.maynard*} *sheffield.ac.uk*

Abstract

News media typically present biased accounts of news stories, and different publications present different angles on the same event. In this research, we investigate how different publications differ in their approach to stories about climate change, by examining the sentiment and topics presented. To understand these attitudes, we find sentiment targets by combining Latent Dirichlet Allocation (LDA) with SentiWordNet, a general sentiment lexicon. Using LDA, we generate topics containing keywords which represent the sentiment targets, and then annotate the data using SentiWordNet before regrouping the articles based on topic similarity. Preliminary analysis identifies clearly different attitudes on the same issue presented in different news sources. Ongoing work is investigating how systematic these attitudes are between different publications, and how these may change over time.

1 Introduction

Editorial decisions in newspaper articles are influenced by diverse forces and ideologies. News publications do not always present unbiased accounts, but typically present frames reflecting opinions and attitudes which can heavily influence the readers' perspectives (Spence and Pidgeon, 2010). Climate change is a controversial issue in which this kind of framing is very apparent. Although bias among different news sources has been discussed previously (Fortuna et al., 2009; Evgenia and van Der Goot, 2008), sentiment analysis has not been commonly applied to newspaper articles for this purpose.

Sentiment analysis is typically implemented on short documents such as Twitter (Pak and Paroubek, 2010; Agarwal et al., 2011) and customer reviews (Pang et al., 2008; Shelke et al., 2017). However, newspaper articles have diverse context length, so their content is much more complicated than other types of sources, especially as these articles are normally cross-domain. A variety of topics might be discussed in the context of a particular climate change issue. Thus, we need to understand what the target of the opinion is in each case, i.e. which aspect of climate change the opinion is about. For instance, using the methods described in this work, we found in reports about the IPCC 2008 (Intergovernmental Panel on Climate Change) that The Independent talked about carbon dioxide emission, but The Guardian concentrated on issues of rising sea levels.

Furthermore, unlike with short documents where one can just find a single sentiment for that document, in order to understand the overall opinion in articles about climate change, we need to look at each opinion and its target separately, as multiple targets may be addressed in a single article. Additionally, even when reporting on the same event and topic, different newspaper sources will have diverse focuses. However, unlike with tweets or customer reviews, newspaper articles must give at least some semblance of objectivity, and often refrain from using explicit positive or negative vocabulary.

In this paper, we examine a set of articles about climate change in four UK broadsheets during the last decade. It is impractical to manually identify topics and analyse all the opinions about them in this large set. We therefore propose a topic modelling method to generate topics using Latent Dirichlet Allocation (LDA), and then cluster the articles into groups with similar topics. Then we perform sentiment analysis on each cluster, in or-

Proceedings of the 2017 EMNLP Workshop on Natural Language Processing meets Journalism, pages 25–30
Copenhagen, Denmark, September 7, 2017.

der to investigate the opinions, how they differ in the 4 sources, and how they may have changed over time.

2 Related Work

Research on sentiment analysis for news articles is not entirely new (Yi et al., 2003; Wilson et al., 2005). Henley et al. (2002) analysed violence-related reports in different newspapers and found that there is a significant difference between the manner of reporting the same violence-related issues. They also found newspaper sentiments reflecting the corresponding ideologies of the editors. However, they applied their content analysis on a limited number of articles, so that the vocabulary for the analysis was also small and strict. Wiebe et al. (2004) applied a classification task for detecting subjectivity and objectivity in newspaper articles. Their work depended on several newspaper datasets which were manually labelled.

Sentiment analysis has been more commonly implemented on newspaper titles. Strapparava and Mihalcea (2007) automatically classified titles with a valence indication, while Burget et al. (2011) proposed a method that classified 6 emotions in Czech newspapers based on their headlines. Burscher et al. (2016) proposed selection and baseline approaches to analyse sentiments in headlines and entire articles respectively, with clustering performed by combining K-means cluster analysis and sentiment analysis. Others have analysed the quotations in newspaper articles. Balahur et al. (2009) extracted annotated quotations from Europe Media Monitor (EMM), and classified them into positive and negative classes using several sentiment lexicons and a Support Vector Machine (SVM) classifier. Both quotations and headlines are short pieces of text, which means that the sentiment analysis is less noisy, and also that the source and target of the sentiment could easily be identified. However, those short pieces of text could not always reveal the insights of news, missing much useful information.

LDA is a generative probabilistic model which has been used to extract abstract topics from documents. It investigates the hidden semantic structure from large amounts of text without requiring manual coding, thus reducing time and cost (Blei et al., 2003). Feuerriegel et al. (2016) applied LDA to extract 40 topics from German financial newspaper articles and found that some topics have an important effect on the stock price market. Xu and Raschid (2016) also developed two probabilistic financial community models to extract topics from financial contracts. However, the implementation of LDA on newspaper articles is less known.

3 Method

3.1 Data

The data for our experiment consists of 11,720 newspaper articles collected from 4 UK broadsheets – *The Guardian, The Times, The Telegraph* and *The Independent* – between 2007 and 2016. These articles were extracted from LexisNexis by searching all four sources for those containing the keywords "Climate Change" at least 3 times in total.

3.2 Pre-processing

In order to identify the topics that can best represent events and issues with respect to climate change, we use a part of speech tagger to annotate all the words, and only keep the nouns for the LDA model. For the sentiment analysis, all words are included.

3.3 LDA model

Typically, the number of topics in the LDA model is determined by computing the log-likelihood or perplexity. However, Bigelow (2002) has shown that predictive likelihood (or equivalently, perplexity) and human judgment are often not correlated, and even sometimes slightly anti-correlated. In this paper, we therefore treat the topics as clusters, and apply the Silhouette Coefficient instead. This method has been previously used for finding the optimal number of topics (Panichella et al., 2013; Ma et al., 2016), and is suitable for our LDA approach, since LDA is fully unsupervised. Nevertheless, in future work, it may be worth evaluating some probability measures such as log-likelihood and perplexity, and comparing the performance using these methods.

$$Sil = \frac{b - a}{max(a, b)} \tag{1}$$

where *a* is the mean distance between a point and other points in the same cluster, and *b* is the mean distance between a point and other points in the next nearest cluster. In the silhouette analysis (Ma et al., 2016), silhouette coefficients close to 1 indicate that the samples in the cluster are far away

Sources	Topics
The Guardian	copenhagen,world,deal,agreement,summit,president,obama,china,action,treaty
The Times	copenhagen, world, cent, deal, president, summit, agreement, conference, china, year
The Telegraph	world, carbon, copenhagen, summit, deal, cent, agreement, energy, time, president
The Independent	world, carbon, copenhagen, deal, cent, agreement, year, conference, cancun, government

Table 1: Topics in 2009

Topic_ID	Keywords
Topic 1	0.31 food 0.84 land 0.79 world ...
Topic 2	0.53 year 0.98 science 0.03 time ...
Topic 3	0.29 world 0.21 car 0.18 weather...

Table 2: Example of Topic list in The Guardian 2007

Articles	Topic_ID	Distributions
Article 1	1	0.519842
Article 2	12	0.348175
Article 3	7, 12	0.412394, 0.1492813
Article 4	2	0.249132

Table 3: Example of topic-document matrix

from the neighbouring clusters. In contrast, a negative silhouette coefficient means that the samples might have been assigned to the wrong cluster.

In our case, we repeatedly ran the analysis on the entire dataset with a different number of topics (0-30) and added the silhouette value for each number of topics to the plot in Figure 1. We can see that when the number of topics reaches 20, it has the highest silhouette coefficient score which indicates the best clustering result.

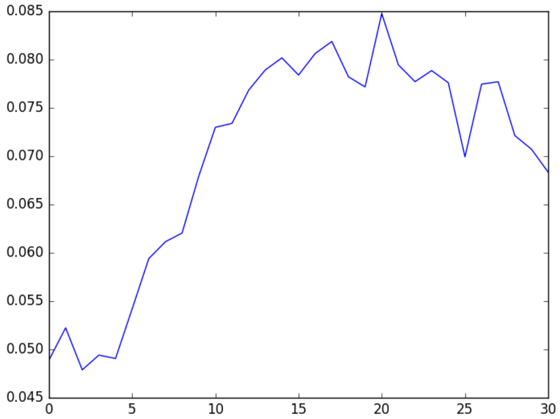


Figure 1: Silhouette analysis for LDA model

Once the number of topics has been determined at 20, the LDA assigns keywords to one of the topics of the news article, based on the probability of the keywords occurring in the topics. This assignment also gives topic representations of all the articles. We repeatedly updated the assignment for 50 iterations to generate both topic distribution in the articles and word distribution in the topics. For each topic in the LDA model, we select the top 10 keywords with their distribution to represent the corresponding topic (see Table 2).

Each article is assigned to a set of topics, and each topic generates a set of keywords based on the vocabulary of the articles. After acquiring the topics from the LDA model, we convert the bag-of-words model into a topic-document matrix, which can be seen as a lower dimensionality matrix (Table 3).

We then select the highest distribution topic among 20 topics from each news article in different news sources.

3.4 Applying SentiWordNet

To automatically annotate the articles with sentiment labels, we use SentiWordNet[1], which contains roughly 155,000 words associated with positive and negative sentiment scores. The keywords in each topic indicate the sentiment targets to be annotated with the corresponding score from SentiWordNet. For each article, the scores for all targets are combined and normalised (to a score between -1 and 1) to deal with the fact that some clusters have more articles than others. The different attitudes of each news source on the same climate change issue can then be analysed once we have a score for each article. For this, we manually check the keywords in the topic lists in each news source in each year, and group those topics containing at least two of the same keywords. Specifically, we analysed every keyword in each topic ID from 2007 to 2016 in each news source, and extract the keywords which occur in each topic. Then we also extract the topic IDs based on those keywords, and group the IDs based on the topics that contain at least two identical keywords. We assume that those news articles have similar or the same topics, as well as sentiment targets, though this also requires verification. We note that

[1] http://sentiwordnet.isti.cnr.it/

Detected Sentences
Positive
China itself defended its crucial role in saving the Copenhagen conference from failure. (The Guardian, 28 Dec, 2009)
Don't panic. Copenhagen really wasn't such a disaster. (The Independent,15 Dec, 2009)
Negative
The move emerged from the chaotic Copenhagen conference on climate change. (The Telegraph, 21 Dec, 2009)
Copenhagen puts nuclear options at risk. (The Times, 23 Dec, 2009)

Table 4: Example sentences with sentiment polarity detected in the four news source in 2009.

the current method of grouping similar topics between news sources manually could introduce human bias. Future work will look at ways to avoid this.

4 Results and Discussion

We compared the 4 news sources by analysing the clusters we identified. For some years, there was no single topic that appeared in the clusters (probably because different newspapers attached different levels of importance to most topics). One example that stands out, however, is the reporting by all 4 broadsheets of the Copenhagen Summit in 2009 (see Table 1). The clusters all contain the keywords "copenhagen" and "agreement", which refer to the Copenhagen Summit explicitly. This feature identified the main topics that also can be seen as the sentiment targets. We utilised this feature to compare the different attitudes toward the same issue (Copenhagen Summit) between four news sources. However, the keywords are mostly different between the sources in other years. For instance, some topics in *The Guardian* and *The Times* have large numbers of keywords such as "gas" and "energy" in 2012, but topics in the *The Telegraph* in that year are associated with the keyword "wind", while *The Independent* has keywords like "government" and "investment".

In Figure 2, we show how sentiment differs between the reports about the Copenhagen Summit in 2009 in the 4 newspapers. Table 4 gives also some examples of positive and negative sentences found. A manual check of a random selection of the relevant articles confirms the general tendency. Most of the articles used some negative words, such as "failures", "collapse", "drastic". However, Figure 2 indicates that the overall sentiment is relatively impartial to positive (the average sentiment score across all sources is 0.15). *The Guardian* is the most positive, while *The Times* is the most negative. We suspect that some of the keywords may be a bit misleading (e.g agreement is typically positive), which might influence the sentiment analysis.

However, there are some clear indications that match the automatic analysis results. While *The Guardian* does have some quite negative reports about the summit, mentioning things like "catastrophic warming", it also tries to focus on the hope aspect ("The talks live. There is climate hope. A bit. Just."). *The Independent* tends also towards the positive, talking about leaders achieving "greater and warmer agreement". The Telegraph, on the other hand, plays more on the fear and alarmist aspect, talking about "drastic action" and "imminent dangerous climate change", although also about positive steps towards the future. The Times, on the other hand, emphasises the role of honesty; although its overall tone is not overwhelmingly negative, it does mention repeatedly the fear and alarmist aspect of climate change and some of the negative points about the summit (for example that Obama will not be there).

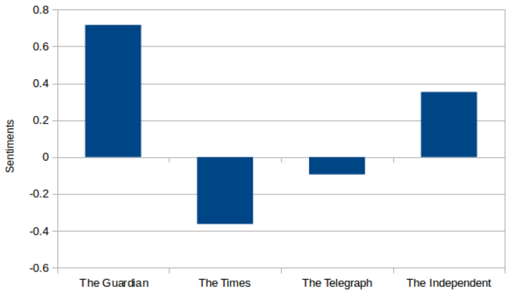


Figure 2: Attitudes of four news sources to the Copenhagen Summit in 2009

In future work, we plan a number of improvements. SentiWordNet is not ideal because it does not cover all the terminology in the specific domain of climate change, nor does it deal with context (see (Maynard and Bontcheva, 2016) for a discussion on these points). We will therefore develop a semi-supervised learning approach, based on a small corpus of manually annotated news articles that we will create, combining lexicon-based and corpus-based methods with co-training, in order to take the best of each. The lexicon-based method will combine LDA with word-embeddings to build a domain-specific lexicon, while the corpus-based method will use a stacked denoising auto-encoder to extract features from news articles. The preliminary results demonstrate the comparison of attitudes between different publications in a single year. However, the attitude towards such climate change topic may change over time. Ongoing work is investigating how the attitudes may change over time between different publications.

5 Conclusion

In this paper, we have described a methodology and a first experiment aimed at understanding the attitudes expressed by different newspapers when reporting about climate change. Traditionally, these kind of analyses have only been carried out manually, and are therefore limited to small case studies. Our aim, however, is to apply such techniques on a large scale, looking at thousands of documents and studying the differences over time, geographic area and newspaper type. While this is only one example about different attitudes to an event, it nevertheless shows a nice case study about how we might use the approach to analyse the different attitudes expressed in the news about the same topic.

Due to the difficulty of annotating news articles manually, and the fact that existing labelled data is rare, an unsupervised approach is more suitable in this case. In contrast to most of the existing sentiment classification approaches, our method is fully unsupervised, which provides more flexibility than other supervised approaches. The preliminary results demonstrate that our method is able to extract similar topics from different publications and to explicitly compare the attitudes expressed by different publications while reporting similar topics.

The methodology is domain-independent and could also be applied to different languages given appropriate lexical resources. Besides the co-training approach mentioned above, there are a number of other ways to extend this work: in particular, we aim to extend the sentiment analysis to consider not just positive and negative attitudes, but also the emotions expressed, and to analyse the effect this might have on readers. The current method also ignored word ordering, so that issues like negation are not considered. We therefore will extend our method to include higher order information in our future experiments.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics, pages 30–38.

Alexandra Balahur, Ralf Steinberger, Erik Van Der Goot, Bruno Pouliquen, and Mijail Kabadjov. 2009. Opinion mining on newspaper quotations. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*. IEEE, volume 3, pages 523–526.

Cindi Bigelow. 2002. Reading the tea leaves. *New England Journal of Entrepreneurship* 5(1):1.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Radim Burget, Jan Karasek, and Zdeněk Smekal. 2011. Recognition of emotions in czech newspaper headlines. *Radioengineering* 20(1):39–47.

Bjorn Burscher, Rens Vliegenthart, and Claes H de Vreese. 2016. Frames beyond words: applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review* 34(5):530–545.

Belyaeva Evgenia and Erik van Der Goot. 2008. News bias of online headlines across languages. *The study of con ict between Russia and Georgia* 73:74.

Stefan Feuerriegel, Antal Ratku, and Dirk Neumann. 2016. Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on*. IEEE, pages 1072–1081.

Blaz Fortuna, Carolina Galleguillos, and Nello Cristianini. 2009. Detection of bias in media outlets with statistical learning methods. *Text Mining* page 27.

Nancy M Henley, Michelle D Miller, Jo Anne Beazley, Diane N Nguyen, Dana Kaminsky, and Robert Sanders. 2002. Frequency and specificity of referents to violence in news reports of anti-gay attacks. *Discourse & Society* 13(1):75–104.

Shutian Ma, Chengzhi Zhang, and Daqing He. 2016. Document representation methods for clustering bilingual documents. *Proceedings of the Association for Information Science and Technology* 53(1):1–10.

Diana Maynard and K. Bontcheva. 2016. Challenges of Evaluating Sentiment Analysis Tools on Social Media. In *Proceedings of LREC 2016*. Portoroz, Slovenia.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*. volume 10.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

Annibale Panichella, Bogdan Dit, Rocco Oliveto, Massimiliano Di Penta, Denys Poshyvanyk, and Andrea De Lucia. 2013. How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms. In *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, pages 522–531.

Nilesh Shelke, Shriniwas Deshpande, and Vilas Thakare. 2017. Domain independent approach for aspect oriented sentiment analysis for product reviews. In *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*. Springer, pages 651–659.

Alexa Spence and Nick Pidgeon. 2010. Framing and communicating climate change: The effects of distance and outcome frame manipulations. *Global Environmental Change* 20(4):656–667.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, pages 70–74.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics* 30(3):277–308.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, pages 347–354.

Zheng Xu and Louiqa Raschid. 2016. Probabilistic financial community models with latent dirichlet allocation for financial supply chains. In *Proceedings of the Second International Workshop on Data Science for Macro-Modeling*. ACM, page 8.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, pages 427–434.

Appendix B

Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network

Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, Diana Maynard
Department of Computer Science
University of Sheffield
Sheffield , UK
{yjiang18,johann.petrak,x.song,
k.bontcheva,d.maynard}@sheffield.ac.uk

Abstract

This paper describes the participation of team "bertha-von-suttner" in the SemEval2019 task 4 Hyperpartisan News Detection task. Our system[1] uses sentence representations from averaged word embeddings generated from the pre-trained ELMo model with Convolutional Neural Networks and Batch Normalization for predicting hyperpartisan news. The final predictions were generated from the averaged predictions of an ensemble of models. With this architecture, our system ranked in first place, based on accuracy, the official scoring metric.

1 Introduction

Hyperpartisan news is typically defined as news which exhibits an extremely biased opinion in favour of one side, or unreasoning allegiance to one party (Potthast et al., 2017). SemEval-2019 Task 4 on "Hyperpartisan News Detection" (Kiesel et al., 2019) is a document-level classification task which requires building a precise and reliable algorithm to automatically discriminate hyperpartisan news from more balanced stories.

One of the major challenges of this task is that the model must have the ability to adapt to a large range of article sizes. In one of the training data sets, the *by-publisher* corpus, the average article length is 796 tokens, but the longest document has 93,714 tokens. Most state-of-the-art neural network approaches for document classification use a token sequence as network input (Kim, 2014; Yin and Schütze, 2016; Conneau et al., 2016). This implies either a high computational cost when a very large maximum sequence length is used to fully represent the longest articles, or alternatively potentially a significant loss of information if the sequence length is restricted to a manageable number of initial tokens from the document.

In this paper, we introduce the **E**LMo **S**entence **R**epresentation **C**onvolutional (ESRC) Network. We first pre-calculate sentence level embeddings as the average of ELMo (Peters et al., 2018) word embeddings for each sentence, and represent the document as a sequence of such sentence embeddings. We then apply a lightweight convolutional Neural Network (CNN), along with Batch Normalization (BN), to learn the document representations and predict the hyperpartisan classification.

Two types of data set have been made available for the task. The *by-publisher* corpus contains 750K articles which were automatically classified based on a categorization of the political bias of the news source. This dataset was split into a training set of 600K articles and a validation set of 150K articles, where all the articles in the validation set originated from sources not in the training set. The second set, *by-article*, contains just 645 articles which were labelled manually. The final evaluation (Potthast et al., 2019) was carried out on a dataset of 628 articles which were also labelled manually.

We created several models based on the two datasets and evaluated them using cross-validation on the *by-article* training set (as the final test set was not available to the participants and it was only available for a maximum of three evaluations). In order to investigate the usefulness of the *by-publisher* training data for training a model that performs well on the manually annotated *by-article* corpus, we experimented with various kinds of pre-training and fine-tuning, and found that any kind of use of the *by-publisher* corpus was actually harmful and decreased the usefulness of the model. A CNN model which used ELMo-based sentence embeddings to represent the article, and was trained on the *by-article* set only,

[1] The code is available at
https://github.com/GateNLP/semeval2019-hyperpartisan-bertha-von-suttner

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 840–844
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

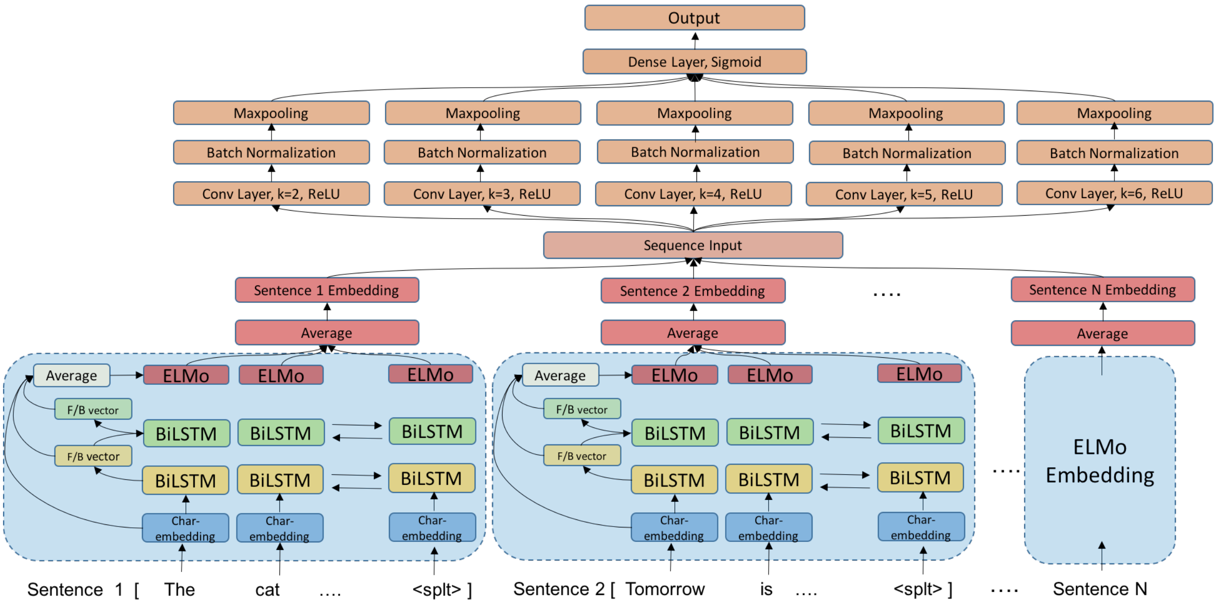


Figure 1: System architecture, *F/B vector* denotes Forward/Backward hidden state from BiLSTM layers.

turned out to outperform all other attempts.

2 System Description

In our model, we represent each article as a sequence of sentence embeddings, where each sentence embedding is calculated as the averaged word embeddings generated from a pre-trained ELMO model. The network consists of 5 parallel convolutional layers with kernel sizes 2,3,4,5,6 and 512 output features, each followed by a ReLU non-linearity, batch normalization, and max-pooling. All the results of the max-pooling layers are combined and go through a final fully connected layer with a sigmoid activation function for the final binary classification. Our model architecture is shown in Figure 1.

2.1 Data

The maximum, mean, and minimum numbers of tokens in the *by-article* corpus are: 6470, 666, 19 respectively, and in the *by-publisher* are: 93714, 796, 10 respectively. This makes it impractical to directly use word level representations as the input for our models. As a simple and easy to calculate compromise between representing the details of the article and as much of a longer article as possible, we represent the article as a sequence of sentence embeddings which are calculated as the average of the word embeddings of a sentence. This can be done using any pre-trained word embeddings and does not require a large training set for training or pre-training, so can be easily applied to even the small *by-article* corpus. To form the input sequence for our network, a maximum of the 200 initial tokens per sentence was used for each sentence embedding and a maximum of 200 sentences was used per article. The title of the article was used as the first article sentence for each document.

2.2 Preprocessing

Our model is character-based, which enabled us to only perform minimal pre-processing. We extract the title and article text from the original XML representation. All the original HTML paragraphs in the text cause a sentence break; the remaining text paragraphs have been split into sentences using `Spacy`. The original case of the text was maintained.

Whitespace is normalized to a single space between tokens; numbers are replaced by a special number token; and all punctuation and other special characters are preserved as input to the pre-trained ELMo model.

2.3 Deep Contextualized Word Representation

Traditionally, the input to CNNs is a set of pre-trained word vectors such as Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), or Fasttext (Bojanowski et al., 2017). In our model, we use the AllenNLP library to generate ELMo embeddings, in which the word representation is learned from character-based units as well as contextual information from the news articles. These character-based word representations allow our model to pick up on morphological features that word-level embeddings could miss, and a valid word representation can be formed even for out-of-vocabulary words. Furthermore, ELMo uses two bi-directional LSTM (Gers et al., 1999) layers to learn the contextual information from the text, which makes it capable of disambiguating the same word into different representations based on its context.

We use the original[2] pre-trained ELMo model to output three vectors for each word. Each vector corresponds to a layer output from the ELMo pre-trained model. Then, we take the average of all three vectors to form the final word vector, and compute the sentence vector by averaging the word vectors in the sentence.

2.4 Convolutional Layers

We combine 5 convolutional layers for different kernel sizes. Each layer is then followed by a non-linear activation function ReLU.

2.5 Batch Normalization

Batch Normalization (BN) is a method for reducing internal covariate shift in neural networks (Ioffe and Szegedy, 2015). BN normalizes the input distribution by subtracting the batch mean and dividing by the batch standard deviation, so that the ranges of input distribution between each layer stay the same. This allows the model to have a higher learning rate, so that the training speed is accelerated. It also reduces overfitting by decreasing the dependence of weight initialization between each layer. The original paper suggested that BN should be applied before the activation layer, but we apply it after the activation layer, after observing better performance in our model this way round. We also applied weighted moving-mean and moving-variance to avoid updating the mean and variance so aggressively in the mini-batch during training time.

2.6 Fully Connected Layer

We perform max-pooling on the output of the batch-normalization layers. Then the outputs of the max-pooling for all convolution layers are combined to form the input to a fully connected layer, which maps to a single output, followed by the Sigmoid function for the binary classification task.

3 Experiments and Results

The generated ELMo embedding contains three vectors for each word, where each vector corresponds to one of the output layers from the pre-trained model. We average the three vectors to generate word representations which contain morphological and contextual information, and compute the sentence vectors by averaging all the word vectors in each sentence. We take a maximum of 200 words for each sentence and a maximum of 200 sentences for each article. If a document has fewer than 200 sentences, we pad the number of sentences out to 200.

Our models are built by using the Keras library with a Tensorflow backend. All the results are shown in Table 1. The table shows for each model the accuracy obtained on the *by-article* training set, and for the submitted models, the *by-publisher* test set, and the hidden *by-article* test set (which unlike the other two, was not available to participants).

In order to investigate the correlation between the two datasets, we first built the `ESRC-publisher` model which is trained on a randomly selected 100K out of the 750K articles from the *by-publisher* corpus, as it is impractical to generate ELMo embeddings for the entire corpus. We also fine-tuned the `ESRC-publisher` model based on the *by-article* set to obtain the `ESRC-publisher-article` model by freezing the weights of all but the last layer of the model. Finally we trained the `ESRC-article` model only on the *by-article* set, one version without and one version (`ESRC-article-BN`) with the additional batch normalization (BN) layer. The accuracy for the `ERC-publisher` model is from evaluating on the whole `by-article` training set, while all other evaluations on the `by-article` training set were carried out using a 10-fold cross validation. However, because of the very limited size of that corpus, the evaluation part of each fold was also used for early stopping and model selection within each fold.

For the evaluation on the hidden test set, we selected the best three models from the 10-folds, according to the accuracy on the evaluation set of

[2] `elmo_2x4096_512_2048cnn_2xhighway`

each fold to form an averaged ensemble model, ESRC-article-BN-Ens.

For comparison, the table also shows the results for an earlier version of the model, GloVe-article, which used GloVe word embeddings (6 billion words, 300 dimensional) to represent up to the first 400 words of the article and did not use batch normalization.

Models	**By-Article Training**
GloVe-article	0.7963
ESRC-publisher	0.5643
ESRC-publisher-article	0.8189
ESRC-article	0.8182
ESRC-article-BN	0.8387
ESRC-article-BN-Ens	**0.8404**
Submitted Models	**By-Article Test**
GloVe-article	0.7659
ESRC-article-BN-Ens	0.8216
Submitted Models	**By-Publisher Test**
GloVe-article	0.6435
ESRC-article-BN-Ens	0.5947

Table 1: System comparison (accuracy).

The parameters in our models are as follows: we used 5 convolutional layers with kernel sizes ($k = 2, 3, 4, 5, 6$) and 512 output features. The momentum in the batch normalization is set to 0.7.[3] We used the default Adam algorithm as the optimizer, and Binary Cross-Entropy as the loss function. The batch size was set to 32 and the fixed number of epochs used was 30. The final best model after 30 epochs was used.

4 Discussion and Conclusion

The ESRC-publisher model performs extremely badly on the *by-article* evaluation data. Even fine-tuning the ESRC-publisher model on the *by-article* corpus produces models which perform worse than a model that is trained only on the *by-article* data. This confirms results from earlier experiments with simpler models that any use of the *by-publisher* data only hurts the model. We assume that the algorithm used for assigning the labels to this dataset just does not reflect any information about hyperpartisan articles sufficiently to be helpful. For this reason, the GloVe-article model also outperforms the ESRC-article-BN-Ens model on the by-publisher dataset.

A quick manual inspection of the data showed that the source of an article is insufficient by far to identify articles as hyperpartisan or not. It would be interesting to know how the algorithm used for creating the *by-publisher* corpus actually performs on the *by-article* corpus. To get maximum performance on the *by-article* dataset, we therefore decided to completely ignore the *by-publisher* data for our final model. The use of BN also showed significant improvement.

Since we use a CNN with a comparatively large number of parameters in relation to the size of the training set which is rather small, we expect significant variance in the generated models and therefore use the average of an ensemble of several models for the final predictions.

5 Acknowledgements

Research partially supported by a Grantham Centre for Sustainable Future Scholarship, a Google Faculty Research Award 2017, and projects funded by the European Commissions Horizon 2020 research and innovation programme under grant agreements No. 654024 SoBigData and No. 825297 WeVerify.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval 2019)*. Association for Computational Linguistics.

[3] This was determined by exploring values from 0.1 to 0.9 at an earlier stage of the experiments and kept, so it may not be the optimal value.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. 2019. TIRA Integrated Research Architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Wenpeng Yin and Hinrich Schütze. 2016. Multichannel variable-size convolution for sentence classification. *arXiv preprint arXiv:1603.04513*.

Appendix C

Comparing Topic-Aware Neural Networks for Bias Detection of News

Comparing Topic-Aware Neural Networks for Bias Detection of News

Ye Jiang[1], Yimin Wang[2], Xingyi Song[1] Diana Maynard[1]
Department of Computer Science[1]
School of Mathematics and Statistics[2]
University of Sheffield
Sheffield , UK
{yjiang18, yimin.wang, x.song, d.maynard}@sheffield.ac.uk

Abstract. The commercial pressure on media has increasingly dominated the institutional rules of news media, and consequently, more and more sensational and dramatized frames and biases are in evidence in newspaper articles. Increased bias in the news media, which can result in misunderstanding and misuse of facts, leads to polarized opinions which can heavily influence the perspectives of the reader. This paper investigates learning models for detecting bias in the news. First, we look at incorporating into the models Latent Dirichlet Allocation (LDA) distributions which could enrich the feature space by adding word co-occurrence distribution and local topic probability in each document. In our proposed models, the LDA distributions are regarded as additive features on the sentence level and document level respectively. Second, we compare the performance of different popular neural network architectures incorporating these LDA distributions on a hyperpartisan newspaper article detection task. Preliminary experiment results show that the hierarchical models benefit more than non-hierarchical models when incorporating LDA features, and the former also outperform the latter.

1 INTRODUCTION

News media typically present biased accounts of news stories, and different ideologies might be presented by different news publications. Detecting bias in the news articles is essential to journalists and researchers for understanding how the presented news stories reflect opinions and attitudes which can heavily influence the readers' perspectives [30]. A growing number of people are consuming biased news, since the hyperpartisan [28] framing style, which exhibits extreme bias, is particularly prone to widespread dissemination on social media. This kind of content has also been identified as a source of increased polarization among the public [23], and consequently leads to further biases in selecting content and in the overall tone of news reporting [16]. Such bias in the news media tends to result in misunderstanding and misuse of facts. Not only is this a factor in swaying individuals' voting preferences [10], but has also even led to ethnic violence [25].

Traditionally, methods such as Latent Semantic Analysis [6], probabilistic Latent Semantic Analysis (pLSA)[11] and Latent Dirichlet Allocation (LDA) [2] have been implemented to infer the semantic meaning of documents through a set of topic representations. Such representations convert text into vector representation which make it feasible for machines to "understand" the semantics of text for tasks such as document summarization [31], document classification [21] and clustering [13]. Methods based on Bag-of-Words (BoW) are frequently used to calculate the statistical features in the document collection. These transform the text data into numeric data that enables a large set of documents to be automatically structured, explored, grouped or clustered based on the word occurrences. However, such document representations suffer from dimensional sparsity, and BoW-based models ignore the contextual information in the text [40], i.e., the relationship between a target word and its surrounding words.

Recently, neural network-based models, which have been proposed in order to generate low-dimensional vector representations, and which are also able to capture semantic word relationships, have been found to outperform most BoW-based models [22, 35]. For instance, the Continuous Bag of Words (C-BoW) model [24] encodes each word into a fixed length vector representation based on other words surrounding the target word. However, such models suffer from the disadvantage that they do not utilize the word co-occurrence of the entire corpus. Specifically, they only scan the textual information within a local context window, which fails to make use of statistical information of the whole corpus. GloVe [27] attempts to resolve this by implementing both global matrix factorization and local content window-based methods; however, our proposal uses a different approach that combines the global co-occurrence information with semantic features of local content windows. Another problem is that many neural network models [39, 5] ignore the hierarchical features of a document, such as the structural relationship between word and sentence, or sentence and document. In an attempt to resolve these issues, we propose a combination of hierarchical frameworks that capture structural features on both word and sentence level, and also incorporate LDA distributions on each level separately.

In order to evaluate the proposed topic-aware hierarchical document representation, we implement a document classification task based on the publicly accessible dataset from the Hyperpartisan News Detection Task.[1] The documents in this corpus are by nature more challenging for learning models than those typically used for traditional document classification (e.g., IMDB, Amazon reviews) for a number of reasons. First, the documents in the hyperpartisan corpus have widely varying length. This means that either the maximum sequence length must be used to fully represent the longest

[1] https://pan.webis.de/semeval19/semeval19-web/index.html

document, which causes a high computational cost, or alternatively a significant information loss will be incurred if the sequence length is restricted to a manageable number of initial tokens from the document. Second, partisanship is more complex than aspects like sentiment to discover, so the learning models require complex text representation to fully capture the subtle semantics.

We perform an evaluation by comparing different popular neural network architectures, with and without incorporating LDA-based distributions, and also compare these with non-hierarchical structures. The code of the proposed model LDA-HAN[2] is available for replicability. Theoretically, the models incorporating LDA distributions should enrich the feature space by adding co-occurrence statistics features and local topic probability distribution on the word and sentence level respectively. Our experimental results demonstrate that the proposed topic-aware document representation outperforms traditional ones, and also that the inclusion of the LDA features has greater impact on the hierarchical representations.

2 RELATED WORK

Traditional BoW-based approaches have often been used to classify newspaper articles. Rubin et al. [29] used a BOW representation with a Support Vector Machine (SVM) to classify satirical news articles. Fortuna et al. [7] also represented news articles in the vector space model by using Term Frequency-Inverse Document Frequency (TF-IDF) weighting, and utilized SVM to identify the bias in describing events in news articles, while Budak et al. [3] used SVM to quantify news bias in a large set of political articles. Meanwhile, LDA has been combined with traditional feature engineering-based methods in many document classification tasks. Wu et al. [36] combined LDA with SVM to classify Chinese news, outperforming the models which generate high-dimensional feature space such as TF-IDF models. Li et al. [18] implemented LDA with a softmax regression to overcome the high dimensional problems of the news text. Kim et al. [14] regarded the document-topic distribution from LDA as a document representation in which both word frequencies and semantic information are considered, to enhance the performance of document classifiers.

Recently, neural network approaches have been combined with LDA for generating document representations. Liu et al. [20] applied LDA to build topic-based word embeddings based on both words and their topics. Xu et al. [37] also implemented LDA to capture topic-based word relationships and then integrated it into distributed word embeddings. Wang and Xu [34] implemented LDA-based text features as input to a deep neural network to detect automobile insurance fraud. Narayan et al. [26] introduced a topic-aware convolutional neural network to generate summaries from online news articles. LDA was used to generate document-topic distributions and word-topic distributions separately, and a CNN was then incorporated to encode and decode the document representations.

However, such approaches generate document representations without considering the characteristics of document structure hierarchically. To address this issue, a Hierarchical Attention Network (HAN) [38] has been previously proposed, which can capture the hierarchical features on both word level and sentence level through a stacked RNN architecture. This outperformed many other baseline models, and indicates that such prior hierarchical information has the potential to enrich document representations, especially when the document sizes are in a wide range.

Hierarchical models have been implemented by many natural language processing (NLP) downstream tasks. Li et al. [17] implemented a hierarchical auto-encoder on both word and sentence level, decoding each representation to reconstruct the original paragraph. Gao et al. [9] constructed a hierarchical convolutional attention model that utilized a combination of self-attention and target-attention. Abreu et al. [1] combined RNN with CNN in a hybrid hierarchical attentional neural network for the document classification task. Zheng et al. [40] compared different hierarchical encoders in documents with differing lengths, and revealed that for document classification, hierarchical frameworks outperform the corresponding neural network models without the hierarchical architecture. They also indicated that the benefits resulting from the hierarchical architecture become more significant as the document length increases. However, these approaches only consider the word embeddings as the input to the encoding layers. Founta et al. [8] utilized a wide variety of available metadata, combining them with word embeddings to enhance the model performance for the task of abusive language detection. Finally, Chen et al. [4] combined word embeddings with WordNet to obtain more relevant occurrences for each sense. Unlike the unified model, which takes different features as inputs to several models independently, our model combines the word embedding with LDA distributions as additive features to the learning model simultaneously.

3 METHODOLOGY

Hierarchical frameworks utilize the document structural features such as the relation between word and sentence, and between sentence and document. Meanwhile, the LDA model generates different distributions which can be used as additional information for encoding document representation. In order to investigate the effectiveness of a learning model which encodes documents hierarchically and incorporates LDA distributions, this paper compares different neural network structures with/without the inclusion of LDA distributions. We first establish three different neural network structures (i.e., CNN, RNN and Transformer) without considering structural features, and then compare these three networks with/without LDA distributions. We also apply two hierarchical models to evaluate the combination of structural features and LDA distributions.

3.1 LDA Distributions

The LDA model generates topic-word distribution and document-topic distribution simultaneously. The former is shared between all documents and contains global word co-occurrence features in the whole corpus, while the latter is the local distribution over the topics for a given document, and is independent of all other documents. These two distributions can be used as additional features in the word level and sentence level encoder layer in the hierarchical frameworks. Each sentence is represented by implementing a specific neural network architecture to encode the combination of word embeddings and transposed topic-word distributions. Similarly, the document is then represented by encoding all sentence representations which are generated from the previous step. Finally, the document representation is concatenated with document-topic distribution as an additional feature to make the final prediction.

3.2 Model Specifications

Let D denote a document consisting of a sequence of sentences (s_1, s_2 , ... , s_m); Meanwhile, let s_i denote a sentence consist-

[2] https://github.com/yjiang18/LDA-HAN

ing of words ($w_{s_i}^1$, $w_{s_i}^2$,..., $w_{s_i}^n$) where $i \in [1, m]$, we embed s_i into a distributional space x = (x_1, x_2, ... , x_n) where $x_j \in \mathbb{R}^k$, $j \in [1, n]$ and k is the dimension of word embedding. Meanwhile, the LDA model generates topic-word distribution, which are transposed as $tw = (tw_1, tw_2, ..., tw_n)$ where $tw_j \in \mathbb{R}^t$ (t denotes number of topics) and the document-topic distribution can be denoted as $dt = (dt_1, dt_2, ..., dt_d)$ where $dt \in \mathbb{R}^{d \times t}$. We train all the models to minimize their cross-entropy error:

$$\ell(\tilde{y}) = \sum_{p=1}^{c} y_p \log(\tilde{y}_p) \tag{1}$$

where $y, \tilde{y}$ are the ground-truth label and predicted label respectively, c denotes number of classes.

3.2.1 *LDA based Non-Hierarchical Models*

Three different network structures are implemented as the encoding layers in the LDA-based non-hierarchical models. Figure 1 depicts the overall model structure. Formally, each document representation is generated from the initial tokens in the document. This is an aggregation of all the word embeddings x to the encoding layer. Meanwhile, the LDA model also takes text input to generate topic-word distribution and document-topic distribution simultaneously. Next, the transposed topic-word distribution tw is concatenated with word embeddings as the input to the encoding layer. The document-topic distribution dt is then concatenated with the generated document representation. Finally, a Fully Connected (FC) layer with softmax activation and Adam optimizer is made for the final classification.

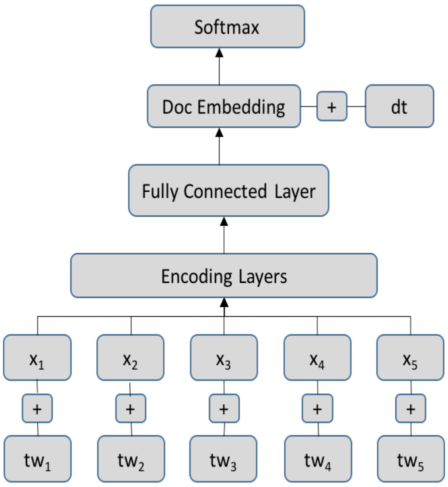


Figure 1. LDA based non-hierarchical models structure

CNN: For a possible variant CNN structure, Kim's implementation [15] is adopted as the baseline CNN model. It consists of 128 filters and 3 different convolutional filter sizes $h \in$ [2,3,4] with ReLU activation, with each convolutional layer followed by a max-pooling layer. The results from the max-pooling layers are concatenated, going through a Fully Connected (FC) layer with 50 hidden units. Formally, the convolutional layer using different filter operators $W_{h,j} \in \mathbb{R}^{h \times k}$ is applied to a window of h words to produce a new feature c_j^h at the word level:

$$c_j^h = \mathit{ReLU}\left((x_{j:j+h-1} \oplus tw_{j:j+h-1}) \circ W_{h,j} + b_{h,j}\right) \tag{2}$$

where the notation $\circ$ and $\oplus$ denote element-wise multiplication and concatenation respectively, ReLU denotes the nonlinear function, $b_{h,j}$ is a bias term. Then, the max-over-time pooling function is used to capture the most important feature $\tilde{c}_j^h$:

$$\tilde{c}_j^h = Max\left(c_j^h\right) \tag{3}$$

The final feature maps are formed by concatenating all $c_j = (\tilde{c}_j^1, \tilde{c}_j^2, ..., \tilde{c}_j^h)$, then the document representation d can be generated by a FC layer:

$$d = \mathit{ReLU}\left(c_j \circ W_j + b_j\right) \tag{4}$$

where W_j is a weight matrix and b_j is a bias term. Finally, the document representation d is concatenated with dt to make the final prediction in a softmax layer.

Self-Attentive RNN: We apply self-Attentive LSTM [19] as the baseline RNN model. It consists of two LSTMs with 50 hidden units and a dropout of probability 0.2 in each direction. In addition, the self-attention layer has 100 hidden units for the outputs from LSTM, and is then followed by an FC layer with 32 hidden units and ReLU non-linearity.

Formally, the forward $\overrightarrow{r_n}$ and backward $\overleftarrow{r_n}$ hidden states at the word level can be obtained by using bidirectional LSTM:

$$\overrightarrow{r_n} = \overrightarrow{LSTM}\left(x_{1:n} \oplus tw_{1:n}\right) \tag{5}$$

$$\overleftarrow{r_n} = \overleftarrow{LSTM}\left(x_{1:n} \oplus tw_{1:n}\right) \tag{6}$$

Then the $\overrightarrow{r_n}$ and $\overleftarrow{r_n}$ can be concatenated as $r_n = (\overrightarrow{r_n}; \overleftarrow{r_n})$, thus each document is encoded as $\tilde{r}_n = (r_1, r_2, ..., r_n)$ where $\tilde{r}_n \in \mathbb{R}^{n \times 2u}$ (u is the hidden unit for each unidirectional LSTM), which is then passed to attention mechanism to get annotation matrix α_n:

$$\alpha_n = softmax\left(W_{s2} Tanh(W_{s1} \tilde{r}_n^T)\right) \tag{7}$$

where $W_{s1} \in \mathbb{R}^{p \times 2u}$, $W_{s2} \in \mathbb{R}^{l \times p}$ (p is the number of neuron units, l denotes to use l times attention) are parameters to learn the important components of the document. The annotation matrix $\alpha_n \in \mathbb{R}^{l \times n}$ multiply $\tilde{r}_n$ to compute the l weighted sums to get the final document representation d.

$$d = \sum_n \alpha_n \tilde{r}_n \tag{8}$$

Finally, the document representation d is concatenated with dt to make the prediction in a softmax layer.

Transformer: We implement the encoder part of Transformer [32] to evaluate its performance on the document classification task. We first calculate the Positional Embeddings (PE) with 300 dimensions for the input, and sum the PE with the original word embeddings instead of concatenation. For the multi-head self-attention, we use a total of eight heads, where each head has 16 units. We then take the average of each step of the output sequence from the self-attention layer, followed by an FC layer with 32 hidden units and ReLU non-linearity. Formally, we use the scaled-dot-product attention to compute the most pertinent information to that document:

$$\mathit{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{k}}\right)V \tag{9}$$

where Q, K, V are 'query', 'key' and 'value' embeddings which concatenate word embeddings x_n with word-topic distribution wt_n. Thus, the final document representation can be formed by multihead attention:

$$\begin{aligned} Multihead(Q, K, V) &= [head_1, head_2, ..., head_n] \\ where \quad head_n &= \mathit{Attention}(Q_j, K_j, V_j) \end{aligned} \tag{10}$$

The final output is the concatenation of the outputs from each head, which is then concatenated with dt to make the final prediction in a softmax layer.

3.2.2 LDA based Hierarchical Models

In this section, we utilize two different hierarchical models to investigate the document representation with/without the LDA features. Figure 2 depicts the overall hierarchical framework structure. The hierarchical models take word and sentence representation as inputs at different phases. The word-topic distribution tw is concatenated with word embeddings x, and is aggregated to a sentence representation to the encoding layer. The document representation can then be formed by aggregating all the sentence representations s. The document-topic distribution dt is concatenated with the generated document representation. An FC layer with softmax activation and Adam optimizer is used for the final classification.

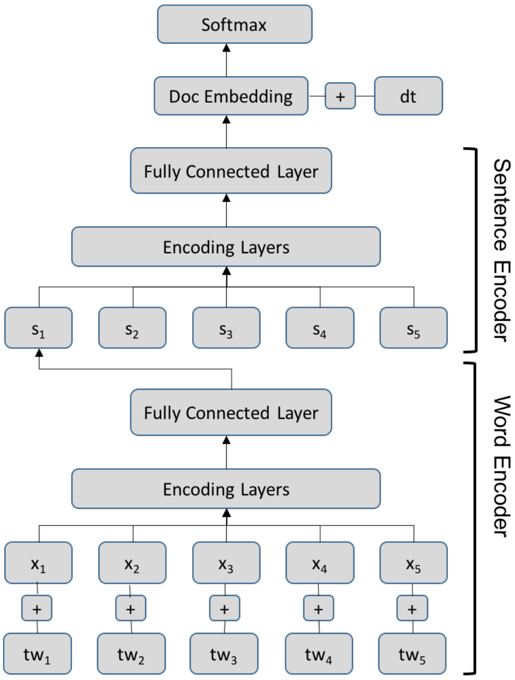


Figure 2. Hierarchical model structure

ESRC: We implemented a similar structure to the ELMo Sentence Representation Convolutional Network (ESRC) [12] for the hierarchical Convolutional framework, but using the pre-trained GloVe embeddings instead of the ELMo embeddings in order to compare them with other hierarchical models. Formally, the word encoder has 128 filters and 7 different convolutional filter sizes $h \in$ [1,2,3,4,5,6,7] with ReLU activation, followed by a batch normalization and a max-pooling layer. The results from the max-pooling layers are concatenated and passed to an FC layer with 32 hidden units and ReLU activation to form a sentence representation. The sentence encoder takes each sentence representation as the input, with the same structure as the word encoder. Similar to Kim's CNN, the encoding convolutional layer using different filter operators $W_{h,j} \in \mathbb{R}^{h \times k}$ is applied to a window of h words to produce a new feature $c_{x_j}^h$ at the word level:

$$c_{x_j}^h = BN\left(ReLU\left(\left(x_{j:j+h-1} \oplus tw_{j:j+h-1}\right) \circ W_{h,j} + b_{h,j}\right)\right) \tag{11}$$

where the notation $\circ$ and $\oplus$ denote the element-wise multiplication and the concatenation respectively, $ReLU$ denotes the nonlinear function, $b_{h,j}$ is a bias term; we also add a batch normalization BN on top of the convolutional layer.

Then, the max-over-time pooling function is used to capture the most important feature $\tilde{c}_{x_j}^h$:

$$\tilde{c}_{x_j}^h = Max\left(c_{x_j}^h\right) \tag{12}$$

The final word-level feature maps are formed by concatenating all $c_{x_j} = (\tilde{c}_{x_j}^1, \tilde{c}_{x_j}^2, ..., \tilde{c}_{x_j}^h)$, then the sentence representation s_i can be generated by an FC layer:

$$s_i = ReLU\left(c_{x_j} \circ W_j + b_j\right) \tag{13}$$

where W_j is a weight matrix and b_j is a bias term. Then, the final document representation d can be obtained similarly: we first obtain sentence-level feature maps $c_{s_i}^h$ by convoluting the sentence sequence using different filter operators, followed by batch normalization:

$$c_{s_i}^h = \left(c_1^h, c_2^h, ..., c_{s_{i:i+h-1}}^h\right) \tag{14}$$

Then, the max pooled features can be obtained:

$$\tilde{c}_{s_i}^h = Max\left(c_{s_i}^h\right) \tag{15}$$

Finally, after concatenating all $\tilde{c}_{s_i}^h$ to obtain c_{s_i} the document representation d can be formed as:

$$d = ReLU\left(c_{s_i} \circ W_i + b_i\right) \tag{16}$$

where W_i is a weight matrix and b_i is a bias term, $ReLU$ is the non-linear function. Finally, the document representation d is concatenated with dt_i to make final predictions in a softmax layer.

HAN: We implement the Hierarchical Attention Network [38] for the hierarchical RNN framework. The word-encoder Bi-LSTM has 100 dimensional hidden units with a dropout of probability 0.2. The sentence encoder has the same structure as the word encoder, except that it has an extra FC layer with 32 hidden units and ReLU nonlinearity.

Formally, the forward $\overrightarrow{r_{x_n}}$ and backward $\overleftarrow{r_{x_n}}$ hidden states at the word level can be obtained by using bi-directional LSTM:

$$\overrightarrow{r_{x_n}} = \overrightarrow{LSTM}\left(x_{1:n} \oplus tw_{1:n}\right) \tag{17}$$

$$\overleftarrow{r_{x_n}} = \overleftarrow{LSTM}\left(x_{1:n} \oplus tw_{1:n}\right) \tag{18}$$

Then the $\overrightarrow{r_{x_n}}$ and $\overleftarrow{r_{x_n}}$ can be concatenated as $r_{x_n} = (\overrightarrow{r_{x_n}}; \overleftarrow{r_{x_n}})$. Together with attention matrix α_n, they are used to calculate the importance of each word. The sentence representation s_m is formed by

$$\alpha_n = softmax(W_{n2}tanh(W_{n1} \circ r_{x_n})) \tag{19}$$

$$s_m = \sum_n \alpha_n r_{x_n} \tag{20}$$

where W_{n1}, W_{n2} denotes the context vector jointly learning the importance of each word in the sentence. Similarly, the document representation d can be also formed by:

$$\overrightarrow{r_{s_m}} = \overrightarrow{LSTM}\left(s_{1:m}\right) \tag{21}$$

$$\overleftarrow{r_{s_m}} = \overleftarrow{LSTM}\,(s_{1:m}) \tag{22}$$

Then the $\overrightarrow{r_{s_m}}$ and $\overleftarrow{r_{s_m}}$ can be concatenated as $r_{s_m} = (\overrightarrow{r_{s_m}}; \overleftarrow{r_{s_m}})$. Together with attention matrix α_m, they are used to calculate the importance of each sentence. The document representation d is formed by

$$\alpha_m = softmax(W_{m2}tanh(W_{m1} \circ r_{s_m})) \tag{23}$$

$$d = \sum_m \alpha_m r_{s_m} \tag{24}$$

where W_{m1}, W_{m2} denotes the context vector jointly learning the importance of each sentence in the document. The document representation d is concatenated with dt to make final predictions in a softmax layer.

4 EXPERIMENTS

We split the dataset into training, evaluation and test sets with a ratio of 8:1:1. We perform 10-fold cross-validation on the training set, then fine-tune and obtain the best performing model based on the evaluation set. The final scores are obtained based on the average of 5 predictions on the test set.

4.1 Dataset

The Hyperpartisan News Detection dataset[3] contains two parts. The *By-Publisher* corpus contains 750K articles which were automatically classified, based on a categorization of the political bias of the news provider. The *By-Articles* corpus contains 1,273 articles which were annotated manually. Although the *By-Publisher* corpus has great potential in training deep learning models due to its significant size, a previous study [12] revealed that there is no significant correlation between the two corpora, in the sense that training a learning model on the *By-Publisher* corpus leads to low performance in the task of predicting partisanship on the *By-Article* corpus. Thus, in this paper all models are only trained on the *By-Article* corpus, as this is more reliable based on its manual annotation assessment [33], and it is also the official ranking corpus for the task. This paper only uses the training set (645 articles) of the *By-Article* corpus, as the rest (628 articles) of the corpus is unavailable to the public (only used for system evaluation). We calculate statistics of `By-Article` as shown in Table 1, and the document length distribution as shown in Figure 3.

Dataset	Hyperpartisan By-Article set
No. of classes	2
No. of documents	645
No. of average sentences/document	31.17
No. of maximum sentences in document	257
No. of average words/sentence	121.13
No. of maximum words in document	5906
No. of average words/document	615.99
No. of words in vocabulary	26135

Table 1. Statistics of dataset

As discussed previously, such a large differentiation in document size makes it impractical to directly use word-level representations as the input, as most news articles have no limitation on sequence length compared to other types of sources (e.g., reviews, tweets, etc). In order to calculate the compromise between representing a summary of the article and as much of its full content as possible, we use the initial 512 tokens to represent each article in the LDA-based non-hierarchical models. For the hierarchical models, we take a maximum of 100 words per sentence, and 30 sentences per document.

4.2 Preprocessing

We extract the title and article text from the original XML file, and represent each article as a sequence of sentences. The text paragraphs are split into sentences, and white spaces are normalized. We used the pre-trained GloVe model[4] to generate word embeddings, and the Gensim LDA model with 425 topics to generate topic-word distribution and document-topic distribution. We use the coherence model to find the optimal number of topics for our LDA model, as shown in Figure 4.

4.3 Results and Discussion

The results, presented in Table 2, show that, on average, the models incorporating LDA distributions outperform the other models. Specifically, the non-hierarchical models have difficulty handling a wide range of document lengths, especially if document sequences are truncated which could potentially cause information loss. Accordingly, the use of hierarchical frameworks, which summarize the importance both on the word level and sentence level features by the corresponding encoders, leads to an improvement in accuracy. Interestingly, the accuracy of the transformer alone is higher than the transformer incorporating LDA distributions, although the transformer models are generally lower than others on accuracy. On the other hand, Attentive-RNN achieves the highest accuracy out of all the non-hierarchical models, especially when it incorporates LDA features. However, the ESRC model gets lower accuracy than most of the non-hierarchical models. The accuracy of this is, however, increased by adding LDA features, and the LDA-ESRC models are better than all the non-hierarchical models. This indicates that the hierarchical frameworks incorporating LDA distributions could improve model performance in terms of accuracy. This is also proved by the LDA-HAN model, which has better accuracy than the HAN model.

Model	Accuracy
Transformer	72.12%
LDA-Transformer	71.56%
CNN	72.95%
LDA-CNN	73.47%
Attentive-RNN	73.63%
LDA-Attentive-RNN	73.75%
ESRC	71.81%
LDA-ESRC	73.69%
HAN	75.69%
LDA-HAN	**76.52%**

Table 2. Performance comparison between models. The best model accuracy is marked in **bold**

.

Although we see that most of the models can be improved by adding LDA features, the hierarchical frameworks can achieve greater improvement from them. The non-hierarchical models can

[3] *https://zenodo.org/record/1489920.XcVDj9Hgrew*

[4] 6 billion words, 300 dimensions

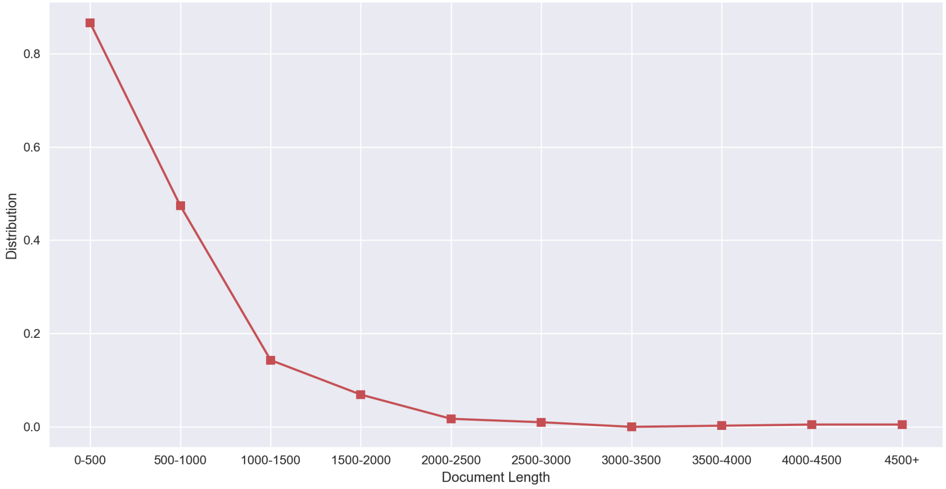


Figure 3. Document size distribution

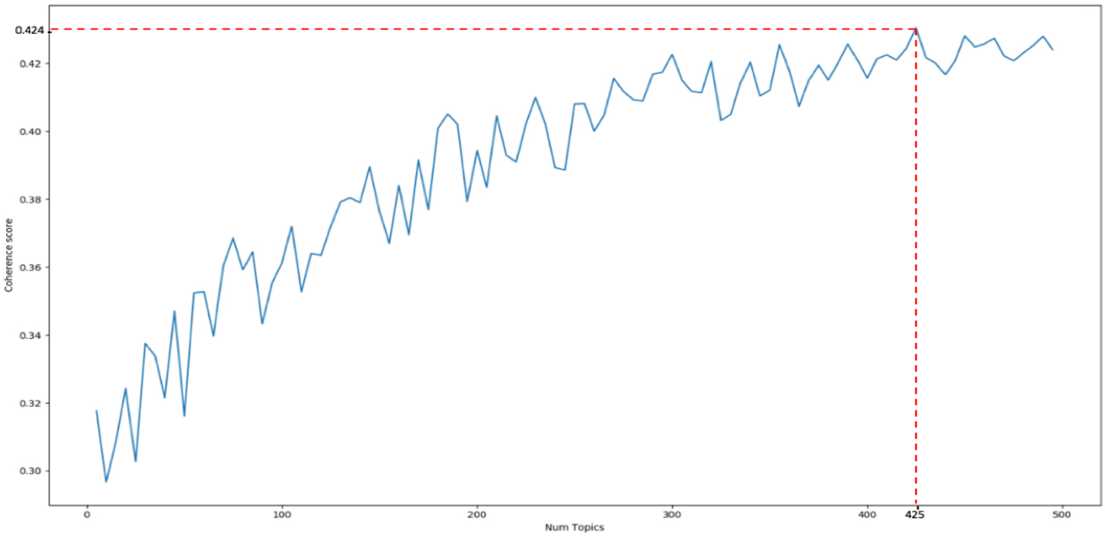


Figure 4. Coherence scores in 500 topics

achieve an improvement of around 0.32% on accuracy, while the hierarchical models can achieve around 1.36% improvement. Specifically, the hierarchical models consider both word-level and sentence-level information separately, and the topic-word distribution enriches the word-level features by adding word occurrence topic distribution through the vocabulary. On the other hand, the document-topic distribution provides local topic distribution, which is independent of all other documents, to increase feature spaces for the final softmax prediction layer, and leads to better accuracy on the document classification task.

5 CONCLUSION

In this paper, we explore the performance of different popular neural network structures with/without incorporating LDA distributions on the recently introduced Hyperpartisan News Detection dataset. This study investigates how the hierarchical models take advantage of the structural features of document to generate a better document representation compared with non-hierarchical models. Meanwhile, the models that include LDA distributions could enrich the feature space by adding global word co-occurrence topic distribution and local document topic probability on word and sentence level respectively.

We first evaluate the non-hierarchical model with/without LDA features. The results demonstrate that most of the non-hierarchical models improved their accuracy when combined with LDA features, except for the Transformer model. On the other hand, most of the hierarchical models achieved better accuracy than non-hierarchical models, and also showed greater improvement when combined with the LDA. This indicates that the hierarchical model has the advantage of handling longer document sequences and reducing information loss by incorporating structural features in the document. Moreover, the benefits resulting from the LDA distributions can be strengthened in the hierarchical models. In conclusion, the combination of hierarchical frameworks and LDA distributions could significantly improve model performance in document classification.

ACKNOWLEDGEMENTS

Research partially supported by a Grantham Centre for Sustainable Future Scholarship, a Google Faculty Research Award 2017, and projects funded by the European Commission's Horizon 2020 research and innovation programme under grant agreements No. 654024 SoBigData and No. 825297 WeVerify.

REFERENCES

[1] J. Abreu, L. Fred, D. Macêdo, and C. Zanchettin. Hierarchical Attentional Hybrid Neural Networks for Document Classification. *arXiv preprint arXiv:1901.06610*, 2019.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[3] C. Budak, S. Goel, and J. M. Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016.

[4] X. Chen, Z. Liu, and M. Sun. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, 2014.

[5] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E17-1104`.

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[7] B. Fortuna, C. Galleguillos, and N. Cristianini. Detection of bias in media outlets with statistical learning methods. In *Text Mining*, pages 57–80. Chapman and Hall/CRC, 2009.

[8] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, pages 105–114. ACM, 2019.

[9] S. Gao, A. Ramanathan, and G. Tourassi. Hierarchical convolutional attention networks for text classification. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2018.

[10] M. Gentzkow. Polarization in 2016. *Toulouse Network for Information Technology Whitepaper*, 2016.

[11] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[12] Y. Jiang, J. Petrak, X. Song, K. Bontcheva, and D. Maynard. Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844, 2019.

[13] M. Keller and S. Bengio. Theme topic mixture model: A graphical model for document representation. In *PASCAL workshop on text mining and understanding*, number CONF, 2004.

[14] D. Kim, D. Seo, S. Cho, and P. Kang. Multi-co-training for document classification using various document representations: Tf–idf, lda, and doc2vec. *Information Sciences*, 477:15–29, 2019.

[15] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[16] N. Landerer. Rethinking the logics: A conceptual framework for the mediatization of politics. *Communication Theory*, 23 (3):239–258, 2013.

[17] J. Li, M.-T. Luong, and D. Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*, 2015.

[18] Z. Li, W. Shang, and M. Yan. News text classification model based on topic model. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–5. IEEE, 2016.

[19] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[20] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical word embeddings. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[21] Y. Lu, Q. Mei, and C. Zhai. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, 14(2):178–203, 2011.

[22] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Won, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. *Ijcai*, 2016.

[23] A. Marwick and R. Lewis. Media manipulation and disinformation online. *New York: Data & Society Research Institute*, 2017.

[24] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[25] M. R. Minar and J. Naher. Violence originated from Facebook: A case study in Bangladesh. *arXiv preprint arXiv:1804.11241*, 2018.

[26] S. Narayan, S. B. Cohen, and M. Lapata. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. *arXiv preprint arXiv:1808.08745*, 2018.

[27] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[28] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv*

preprint arXiv:1702.05638, 2017.

[29] V. Rubin, N. Conroy, Y. Chen, and S. Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17, 2016.

[30] A. Spence and N. Pidgeon. Framing and communicating climate change: The effects of distance and outcome frame manipulations. *Global Environmental Change*, 20(4):656–667, 2010.

[31] J. Steinberger and M. Křišt'an. Lsa-based multi-document summarization. In *Proceedings of 8th International PhD Workshop on Systems and Control*, volume 7, 2007.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[33] E. Vincent and M. Mestre. Crowdsourced measure of news articles bias: Assessing contributors' reliability. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) co-located with the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zürich, Switzerland, July 5, 2018*, pages 1–10, 2018. URL `http://ceur-ws.org/Vol-2276/paper1.pdf`.

[34] Y. Wang and W. Xu. Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, 105:87–95, 2018.

[35] W. Wei, X. Zhang, X. Liu, W. Chen, and T. Wang. pkudblab at SemEval-2016 Task 6 : A Specific Convolutional Neural Network System for Effective Stance Detection. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016. doi: 10.18653/v1/s16-1062.

[36] X. Wu, L. Fang, P. Wang, and N. Yu. Performance of using LDA for Chinese news text classification. In *2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1260–1264. IEEE, 2015.

[37] H. Xu, M. Dong, D. Zhu, A. Kotov, A. I. Carcone, and S. Naar-King. Text classification with topic-based word embedding and convolutional neural networks. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 88–97. ACM, 2016.

[38] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*, 2016.

[39] W. Yin and H. Schütze. Multichannel variable-size convolution for sentence classification. *arXiv preprint arXiv:1603.04513*, 2016.

[40] J. Zheng, F. Cai, W. Chen, C. Feng, and H. Chen. Hierarchical neural representation for document classification. *Cognitive Computation*, 11(2):317–327, 2019.

Appendix D

Hierarchical Contextual Document Representation for Detecting Hyperpartisan News Articles (Submitted)

Hierarchical Contextual Document Representation for Detecting Hyperpartisan News Articles

Anonymous COLING submission

Abstract

The recently released Hyperpartisan News Detection dataset affords great potential for developing methods for the automatic classification of biased news. However, the diversity of document lengths in this dataset produces some challenges. Traditional sentence level representation for such methods, which pads or truncates document sequences to a fixed length, might cause either information loss if the sequence is truncated to a manageable length, or alternatively, high computational cost when padded to the maximum sequence length in the corpus. Meanwhile, traditional learning models encode document representation without considering structural information between sentences and words, especially in documents which contain hundreds of sentences, such as newspaper articles. Also, traditional word embeddings generate a context-free representation which might cause semantic ambiguity. To address these issues, this paper demonstrates how the combination of hierarchical frameworks and recent contextual embeddings could significantly improve the model performance in encoding various sizes of documents. We evaluate this performance on the binary document classification task of hyperpartisan news detection. Preliminary experiment results show that the proposed models outperform many other baseline models.

1 Introduction

News media providers are often accused of exhibiting increasing partisanship in news articles (Martin and Yurukoglu, 2017). This has also been identified as a source of increased polarization among the public (Marwick and Lewis, 2017), something which is generally detrimental to democracy. It can result in misunderstanding and misuse of facts, is a factor in changing individuals' voting preferences (Gentzkow, 2016), and has even led to ethnic violence (Minar and Naher, 2018). This type of news which expresses an extremely one-sided opinion or unreasoning allegiance to one party, has been recently defined as hyperpartisan news (Potthast et al., 2017).

A new dataset has been made openly available in the International Workshop on Semantic Evaluation 2019 (SemEval 2019) task 4: Hyperpartisan News Detection. One of the main challenges of this task is that the learning models must have the ability to adapt to a large range of document sizes. Unlike other types of text resource (e.g. tweets, reviews, etc) which normally are restricted, either formally or informally, to a certain limited document size, news articles are naturally more flexible in terms of the number of paragraphs used to convey one or more points of view. This might cause crucial information loss when partially encoding a document. For instance, in one of the training data sets, the *by-publisher* corpus, the average document length has 796 tokens, but the longest document has 93,714 tokens.

Typically, Bag-of-Words (BoW) (Lan et al., 2009; Joachims, 1998), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and n-grams with Term Frequency-Inverse Document Frequency (TF-IDF) (Wu et al., 2008) are used to generate document representation. However, such models might cause sparsity of document representation and dimensional explosion when the corpus is very large.

Recently, neural network models, which have been proposed to generate low dimensional vectors and are also able to capture semantic word relationships, have been found to outperform most traditional ones (Abreu et al., 2019; Ma et al., 2016; Ruchansky et al., 2017; Zhang et al., 2016). However, most neural network models use token sequences as input (Conneau et al., 2017; Yin and Schütze, 2016). Such models imply either the maximum sequence length is used to fully represent the longest document, which causes a high computational cost, or alternatively a significant information loss if the sequence length is restricted to a manageable number of initial tokens from the documents. Furthermore, such document representation ignores the hierarchical features of a document, such as the structural relationship between word and sentence, or sentence and document.

In an attempt to resolve this issue, this paper proposes a combination of hierarchical frameworks that capture structural features on both word and sentence level, and also generate a context-sensitive document representation. Traditionally, the input to a neural network model is a set of pre-trained word embeddings such as Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), or FastText (Bojanowski et al., 2017). Such word embeddings generate a context-free representation for each word in the vocabulary. For instance, the word "apple" has the same vector representation for the meanings "eat apple" and "the Apple company"; however, the context completely changes the meaning of "apple" in a sentence. For this reason, recent pre-trained language models (e.g. ELMo(Peters et al., 2018) and BERT(Devlin et al., 2018)) utilize bidirectional approaches to guard against context-free issues. Specifically, ELMo and BERT respectively use bidirectional LSTM and Transformer (Vaswani et al., 2017) to learn the contextual information from the text. Additionally, ELMo is learned from character-based units which allow the word embeddings to pick up on morphological features that word-level embeddings could miss, and even enable them to handle out-of-vocabulary problems.

In our proposed model, each sentence is represented by implementing a specific neural network architecture to encode the contextual word embeddings in the sentence. Similarly, the document is represented by encoding all sentence representations which are generated from the previous step. In order to evaluate the hierarchical contextual document representation on various document lengths, we implement a document-level classification task by taking advantage of the document length variety in the Hyperpartisan News dataset, and compare the performance between models. We also compare the contextual word embeddings, which are generated by state-of-the-art language models ELMo and BERT, in different types of hierarchical frameworks. Our experimental results demonstrate that the proposed model outperforms other existing baselines.

2 Related Work

Traditionally, many feature engineering-based approaches have been used to classify documents. Lin (2011) applied Support Vector Machines (SVM) and Naive Bayes (NB) to detect subjectivity on the sentence level classification task. Rubin (2016) used BOW representation with SVM to classify satirical news articles. Fortuna (2009) also represented news articles in the vector space model by using TF-IDF weighting, and utilized SVM to identify the bias in describing the events in news articles, while Budak (2016) used SVM to quantify news bias in a large set of political articles.

Recently, neural network approaches have been used to generate document representations. Iyyer (2014) applied a recursive neural network to identify political ideology evinced by sentence level representation. Ruchansky (2017) used a recurrent neural network to extract temporal text representation to detect fake news. Kim (2014) adopted a convolutional neural network (CNN) to classify documents, and Zhang (2016) implemented multiple sets of word embeddings for generating document representation. Wei (2016) developed a CNN for stance detection in tweets. Ma (2016) applied a recurrent neural network (RNN) to detect rumours from microblogs. Iyyer (2015) used Deep Average Network (DAN), which simply takes the average of word embeddings and passes the averages through one or more feed forward layers, achieving comparable model performance with extremely fast training speed.

However, such approaches generate document representations without considering the characteristics of the document structure hierarchically. To address this issue, Yang (2016) proposed a Hierarchical Attention Network (HAN), which could capture the hierarchical features on both word level and sentence level through a stacked RNN architecture. This outperformed many other baseline models and indicates that such prior hierarchical information has the potential to generate better document representations, especially when the document sizes are in a wide range. Hierarchical models have been implemented by many natural language processing (NLP) downstream tasks. Li (2015) implemented a hierarchical auto-encoder on both word and sentence level, and decode each representation to reconstruct the original paragraph. Gao (2018) constructed a hierarchical convolutional attention model that utilized a combination of self-attention and target-attention. Abreu (2019) combined RNN with CNN in a hybrid hierarchical attentional neural network in the document classification task. Zheng (2019) compared different hierarchical encoders on documents of different lengths, and revealed that hierarchical frameworks outperform the corresponding neural network models without the hierarchical architecture for document classification. They also indicated that the benefits resulting from the hierarchical architecture can be strengthened as the document length increases.

The input to a neural network model is typically a set of pre-trained word embeddings such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) or FastText (Bojanowski et al., 2017). Such word embeddings generate a context-free representation for each word in the vocabulary, which might cause semantic ambiguity in terms of document representation. To address this issue, context-sensitive word embeddings have recently been developed, such as Embeddings from Language Models (ELMo)(Peters et al., 2018) and Bidirectional Encoder Representations from Transformer (BERT) (Devlin et al., 2018), which generate a representation based on its context in the sentence by using Bidirectional LSTM (Gers et al., 1999) and Transformer (Vaswani et al., 2017) respectively, achieving state-of-the-art performance in many downstream NLP tasks. Other previous work combined ELMo embeddings with a light-weight CNN model (Jiang et al., 2019), and the resulting system was ranked first in the SemEval 2019 task 4. Alsentzer (2019) also implemented BERT embeddings on a clinical corpus. Wiedemann (2019) explored the word sense disambiguation of three contextual word embeddings (BERT, ELMo and Flair (Akbik et al., 2018)), and demonstrated that the pre-trained BERT model was able to place polysemic words into distinct 'sense' regions of the embedding space. This paper extends the utilities of contextual word embeddings by incorporating document structural information, and also compares the performance of ELMo embeddings and BERT embeddings in terms of the document level classification task.

3 Methodology

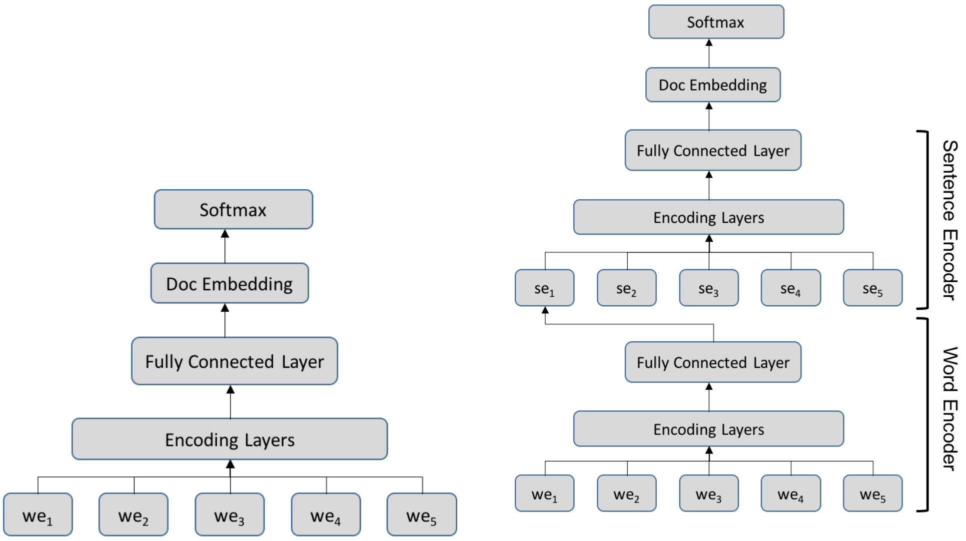
Hierarchical frameworks utilize the document structural features such as the relation between sentences and words. In order to investigate the effectiveness of the learning model encoding document representation hierarchically, this paper compares the different neural network structures with/without incorporating hierarchical frameworks. We first establish three different network structures, without considering structural features as baseline models. Then, we apply hierarchical structures as hierarchical models accordingly on the top of these baseline models. Two different contextual embeddings (ELMo and BERT) are used in both baseline and hierarchical models.

3.1 Baseline Models

Three different network structures are implemented as the baseline models. Figure 1a demonstrates the overall baseline model structure. Formally, each document representation is generated from the initial tokens in the document. This is an aggregation of all the contextual word embeddings we from a specific neural network in the encoding layer. Finally, a Fully Connected (FC) layer with softmax activation and Adam optimizer is made for the final classification.

CNN-base: For a possible variant CNN structure, we implement a light-weight CNN model based on the ELMo Sentence Representation Convolutional Network (ESRC) (Jiang et al., 2019). It consists of 128 filters and 7 different convolutional filter sizes [1,2,3,4,5,6,7] with ReLU activation, followed by a batch normalization layer and a max-pooling layer. The results from max-pooling layers are concatenated and go through an FC layer with 32 hidden units and ReLU non-linearity.

RNN-base: We apply Bidirectional LSTM (Bi-LSTM) as the baseline RNN model. Bidirectional RNN concatenates both forward and backward hidden states, and this characteristic could capture contextual information in each sequence. The Bi-LSTM layer has 100 dimensional hidden units with a dropout probability of 0.2, and is followed by an FC layer with 32 hidden units and ReLU non-linearity.



(a) Baseline model

(b) Hierarchical model structure

Figure 1: Model Architectures

DAN-base: DAN implements neural bag-of-words functions that ignore the sequence order information, but significantly increase training speed and could also achieve comparable model performance. We directly take the average of a fixed number of initial word embeddings to form the document representation, followed by an FC layer with 32 units and ReLU non-linearity. We also use dropout function before contextual embeddings pass to the encoding layer.

3.2 Hierarchical Models

For these, we utilize the hierarchical features on the top of our baseline models. Figure 1b demonstrates the overall hierarchical framework structure. The hierarchical models take word and sentence representation as inputs separately. The contextual word embeddings we are aggregated to a sentence representation using a specific hierarchical neural network. The document representation can then be formed by aggregating all the sentence representations se. Finally, an FC layer with softmax activation and Adam optimizer is made for the final classification.

Let d denote a document consisting of a sequence of sentences (s_1, s_2 , ... , s_m); Meanwhile, let s_i denote a sentence consisting of words ($w^1_{s_i}$, $w^2_{s_i}$,..., $w^n_{s_i}$) where $i \in [1, m]$, we embed s_i into a distributional space x = (x_1, x_2, ... , x_n) where $x_j \in R^k$, $j \in [1, n]$ and k is the dimension of the n-th word embedding in the sentence. We train all the models to minimize the cross-entropy error by:

$$\ell(\tilde{y}) = \sum_{p=1}^{b} y_p \log(\tilde{y}_p) \tag{1}$$

where $y, \tilde{y}$ are the ground-truth label and predicted label respectively, b denotes number of classes.

H-CNN: In the H-CNN model, the word encoder has 128 filters and 7 different convolutional filter sizes $h \in$ [1,2,3,4,5,6,7] with ReLU activation, with each convolutional layer followed by a batch normalization and a max-pooling layer. The results from the max-pooling layers are concatenated to form a sentence representation. The sentence encoder takes each sentence representation as input, with the same structure as the word encoder, except it has an extra FC layer with 32 hidden units and ReLU nonlinearity after the final concatenation. Specifically, the convolutional layer using different filter operators $W_{h,j} \in R^{h \times k}$ is applied to a window of h words to produce a new feature c^h_j at the word level:

$$c^h_{x_j} = BN\left(ReLU\left(x_{j:j+h-1} \circ W_{h,j} + b_{h,j}\right)\right) \tag{2}$$

where the notation $\circ$ denotes element-wise multiplication, $ReLU$ denotes the nonlinear function, $b_{h,j} \in R$ is a bias term, BN denotes batch normalization.

Then, the max-over-time pooling function is used to capture the most important feature $\tilde{c}^{h}_{x_j}$:

$$\tilde{c}^{h}_{x_j} = Max\left(c^{h}_{x_j}\right) \tag{3}$$

The final feature maps c_{x_j} are formed by concatenating all $c_{x_j} = (\tilde{c}^{1}_{x_j}, \tilde{c}^{2}_{x_j}, ..., \tilde{c}^{7}_{x_j})$, then the sentence representation s_i can be generated by an FC layer:

$$s_i = ReLU\left(c_{x_j} \circ W_j + b_j\right) \tag{4}$$

where W_j is a weight matrix and b_j is a bias term. Then, the final document representation d can be obtained similarly: we first obtain the feature maps c^{h}_{i} by convoluting the sentence sequence using different filter operators, and applying batch normalization:

$$c^{h}_{s_i} = \left(c^{h}_{s_1}, c^{h}_{s_2}, ..., c^{h}_{s_i:s_{i+h-1}}\right) \tag{5}$$

Then, the max pooled features can be obtained:

$$\tilde{c}^{h}_{s_i} = Max\left(c^{h}_{s_i}\right) \tag{6}$$

Finally, after concatenating $\tilde{c}^{h}_{s_i}$ to obtain c_{s_i} the document representation d can be formed as d:

d = ReLU($c_{s_i} \circ W_i + b_i$) (7)where W_i is a weight matrix and b_i is a bias term, $ReLU$ is the non-linear function. Finally, the document representation d is formed to make the final prediction in a softmax layer.

H-RNN: We apply two Bi-LSTM encoders to form the H-RNN model. The word-encoder Bi-LSTM has 100 dimensional hidden units with a dropout probability of 0.2, and is followed by batch normalization and an FC layer with 100 hidden units and ReLU activation. The sentence encoder also has the same structure as the word encoder, except it has an extra FC layer with 32 hidden units and ReLU non-linearity.

Formally, the forward $\overrightarrow{r_{x_n}}$ and backward $\overleftarrow{r_{x_n}}$ hidden states at the word level can be obtained by using bidirectional LSTM:

$$\overrightarrow{r_{x_n}} = \overrightarrow{LSTM}\left(x_{1:n}\right) \tag{8}$$

$$\overleftarrow{r_{x_n}} = \overleftarrow{LSTM}\left(x_{1:n}\right) \tag{9}$$

Then the $\overrightarrow{r_{x_n}}$ and $\overleftarrow{r_{x_n}}$ can be concatenated as $r_{x_n} = (\overrightarrow{r_{x_n}}; \overleftarrow{r_{x_n}})$and pass to ReLU non-linear function to form the sentence representation s_m:

$$s_m = ReLU\left(r_{x_n} \circ W_n + b_n\right) \tag{10}$$

where W_n denotes a weight matrix and b_n denotes a bias term. Similarly, the sentence level hidden states can be also formed by:

$$\overrightarrow{r_{s_m}} = \overrightarrow{LSTM}\left(s_{1:m}\right) \tag{11}$$

$$\overleftarrow{r_{s_m}} = \overleftarrow{LSTM}\left(s_{1:m}\right) \tag{12}$$

Then the $\overrightarrow{r_{s_m}}$ and $\overleftarrow{r_{s_m}}$ can be concatenated as $r_{s_m} = (\overrightarrow{r_{s_m}}; \overleftarrow{r_{s_m}})$ and pass to ReLU non-linear function to form the document representation d:

$$d = ReLU\left(r_{s_m} \circ W_m + b_m\right) \tag{13}$$

where W_m denotes a weight matrix and b_m denotes a bias term. Finally, the document representation d is formed to make the final prediction by a softmax layer.

H-DAN: Similar to other hierarchical models, the word encoder takes word embeddings as the input, and then takes the average of the word level representation to form the sentence representation. The sentence encoder then takes the average of the sentence representations to form the document representation. Finally, the document representation is passed through an FC layer that contains 32 units and ReLU non-linearity.

Specifically, the averaged sentence embedding $\tilde{s}$ can be obtained simply by averaging each word embedding x_j in a sentence:

$$\tilde{s} = 1/|n| \sum_{j=1}^{n} x_j^l \quad (l \in [1, k] \text{ and } \tilde{s}, x_j \in R^k) \tag{14}$$

Then, it is passed to a non-linear function $ReLU$ to obtain the final representation s_i:

$$s_i = ReLU(\tilde{s} \circ W_j + b_j) \tag{15}$$

where s_i denotes the sentence representation, W_j denotes a weight matrix and b_j denotes a bias term. Similarly, the document representation can be formed by:

$$\tilde{d} = 1/|i| \sum_{m=1}^{i} s_i^l \quad (l \in [1, k] \text{ and } \tilde{d}, s_i \in R^k) \tag{16}$$

Then, it is passed to a non-linear function $ReLU$ to obtain the final representation:

$$d = f(\tilde{d} \circ W_i + b_i) \tag{17}$$

where W_i denotes a weight matrix and b_i denotes a bias term. Finally, the document representation is passed through an FC layer that contains 32 units and ReLU non-linearity.

3.3 Embedding Generation

The pre-trained BERT[1] and ELMo[2] models are used to generate contextual word embeddings for baseline and hierarchical models. For generating ELMo embeddings, we use the official pre-trained ELMo model from AllenNLP. The original ELMo pre-trained model generates three vectors for each word, where each vector corresponds to a layer output from the ELMo pre-trained language model. The first layer corresponds to the context-insensitive token representation, followed by the two LSTM layers. Then, we take the average of all three vectors to form the final word vector. Specifically, the word representation is learned from character-based units as well as contextual information from the news articles. These character-based word representations allow it to pick up on morphological features that word-level embeddings could miss, and a valid word representation can be formed even for out-of-vocabulary words. Furthermore, ELMo uses two bi-directional LSTM (Gers et al., 1999) layers to learn the contextual information from the text, which makes it capable of disambiguating the same word into different representations based on its context. We implement bert-as-service (Xiao, 2018) to generate BERT embeddings. This is a sentence encoding service based on Google BERT and ZeroMQ, which allows the mapping of a variable-length sentence to a fixed-length vector.

4 Experiments

We perform 10-fold cross validation on the dataset, The final scores are obtained based on the averaged predictions on each fold.

[1] *BERT-Base, Uncased*

[2] `elmo_2x4096_512_2048cnn_2xhighway`

Dataset	Hyperpartisan By-Article set
No. of classes	2
No. of documents	645
No. of average sentences/document	31.17
No. of maximum sentences in document	257
No. of average words/sentence	121.13
No. of maximum words in document	5906
No. of average words/document	615.99
No. of words in vocabulary	26135

Table 1: Statistics of dataset

4.1 Dataset

The Hyperpartisan News Detection dataset[3] contains two parts. The *By-Publisher* corpus contains 750K articles which were automatically classified, based on a categorization of the political bias of the news provider. The *By-Articles* corpus contains 1,273 articles which were annotated manually. Although the *By-Publisher* corpus has great potential in training deep learning models due to its significant size, a previous study (Jiang et al., 2019) revealed that there is no significant correlation between the two corpora, in the sense that training a learning model on the *By-Publisher* corpus leads to low performance in the task of predicting partisanship on the *By-Article* corpus. Thus, in this paper all models are only trained on the *By-Article* corpus, as this is more reliable based on its manual annotation assessment (Vincent and Mestre, 2018), and it is also the official ranking corpus for the task. This paper only uses the training set (645 articles) of the *By-Article* corpus, as the rest (628 articles) of the corpus is unavailable to the public (only used for system evaluation).

We calculate the statistics of `By-Article` as shown in Table 1. As discussed previously, such a large differentiation in document size makes it impractical to directly use word level representations as the input, as most news articles have no limitation on sequence length compared to other types of sources (e.g. reviews, tweets, etc). In order to calculate the compromise between representing a summary of the article and as much of its full content as possible, we use the initial 512 (i.e. the maximum sequence length which can be taken from the pre-trained BERT model) tokens to represent each article in the baseline models. For the hierarchical models, we take a maximum of 100 words per sentence, and 30 sentences per document.

4.2 Preprocessing

We extract the title and article text from the original XML file, and represent each article as a sequence of sentences. The text paragraphs are split into sentences and white spaces are normalized. As the ELMo pre-trained model is character-based, this enables us to only perform minimal pre-processing. In terms of generating BERT embeddings, the original text is lower cased, and punctuation is removed.

4.3 Results and Discussion

The results, presented in Table 2, show that, on average, the hierarchical models outperform the baseline models in terms of accuracy. Specifically, the baseline models have difficulty handling a wide range of document lengths, especially if document sequences are truncated which could potentially cause information loss. Accordingly, the use of hierarchical models, which summarize the importance both on the word level and sentence level features by the corresponding encoders, leads to an improvement in accuracy.

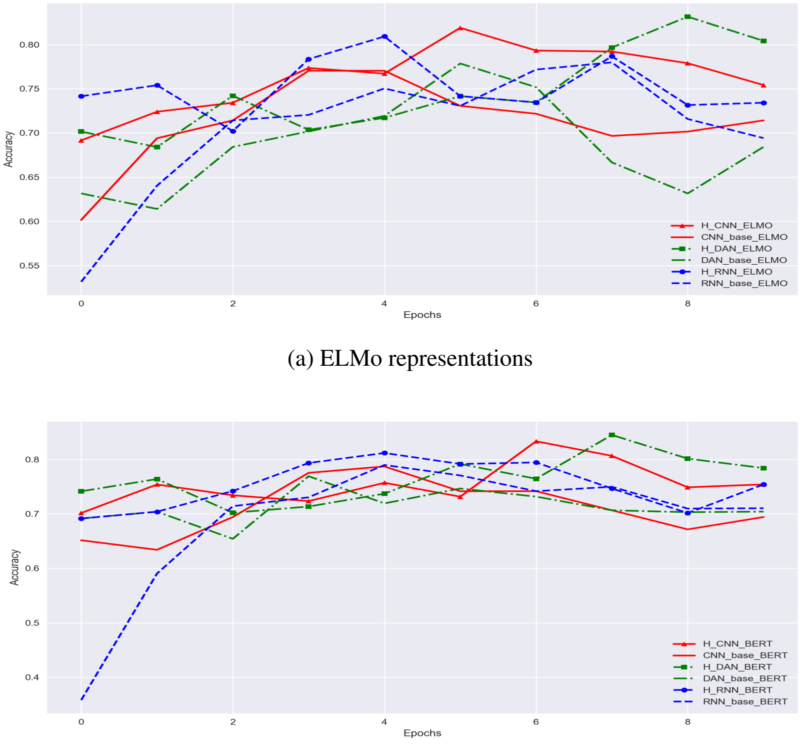
Interestingly, the accuracy of DAN models (i.e. DAN-base and H-DAN) is higher than that of their neural network structures (i.e. RNN-base, CNN-base, H-RNN and H-CNN). This indicates that simply taking the average is better than other architectures (i.e. RNN, CNN). This might be because the small sized data set is causing overfitting in either RNN or CNN models. Although DAN utilizes the unordered

[3] *https://pan.webis.de/semeval19/semeval19-web*

Models	Accuracy (std.)	Precision (std.)	Recall (std.)	F1 (std.)
RNN-base-ELMo	.7785 (.0505)	.8082 (.0385)	.5919 (.0865)	.6833 (.0710)
RNN-base-BERT	.7692 (.0312)	.7731 (.0281)	.6219 (.0789)	.6893 (.0601)
CNN-base-ELMo	.7704 (.0323)	.7846 (.0592)	.6137 (.0577)	.6887 (.0387)
CNN-base-BERT	.7871 (.0358)	.7946 (.0611)	.5937 (.0549)	.6796 (.0379)
DAN-base-ELMo	.7798 (.0421)	.7831 (.0212)	.6134 (.0800)	.6879 (.0381)
DAN-base-BERT	.7898 (.0371)	.7979 (.0303)	.6040 (.0873)	.6875 (.0571)
H-CNN-ELMo	.8189 (.0471)	.7029 (.0281)	.7857 (.1024)	.7420 (.0895)
H-CNN-BERT	.8334 (.0538)	.7320 (.0329)	.7657 (.1480)	.7485 (.0941)
H-RNN-ELMo	.8091 (.0987)	.7534 (.2193)	.7762 (.0461)	.7646 (.0783)
H-RNN-BERT	.8119 (.1292)	.7743 (.3513)	**.8262 (.3361)**	**.7994** (.3633)
H-DAN-ELMo	.8315 (.0992)	**.8401 (.0031)**	.7239(.0913)	.7776(.0531)
H-DAN-BERT	**.8450 (.0482)**	.8338 (.0431)	.7677(.1096)	**.7993(.0682)**

Table 2: Performance comparison between models. Best values are marked in **bold**, standard deviations in parentheses

functions, the positional information is kept in the word representation when ELMo and BERT are generating word level embeddings based on the context of the word. Such neural network architectures still have the potential to outperform methods which take the average by hyperparameter tuning and adding training samples. Additionally, all the baseline models achieve significantly lower recall scores than hierarchical models. This could be explained by the possibility of the baseline models missing matching instances in the training, as the truncated sequence length cannot fully represent features in the document. Also, BERT models are relatively better than ELMo models on average in terms of accuracy.



(a) ELMo representations

(b) BERT representations

Figure 2: Classification accuracy of the training models

The accuracy of H-RNN, H-CNN and H-DAN shows around 5% improvement compared to the RNN-base, CNN-base and DAN-base respectively. However, the training speed of DAN is much faster than others in both base and hierarchical structures, while obtaining comparable performance.

Figures 2a and 2b demonstrate the accuracy of ELMo and BERT models respectively. Generally, most of the models converged in the first 10 epochs. The baseline models converge quicker than the hierarchical models generally. Specifically, the accuracy reaches its peak at around the fourth epoch in the baseline models. However, the accuracy of hierarchical models mostly converges after the sixth epoch. This is expected since the hierarchical model has its ability to encode more sequence representation, so the hierarchical models might take longer for converging.

5 Conclusion

In this paper, we explore the performance of hierarchical models with context-sensitive embeddings on the recently introduced Hyperpartisan News Detection dataset. Specifically, we first evaluate the hierarchical framework compared with the baseline models. The results demonstrate that the hierarchical model has the advantage of handling longer document sequences and reducing information loss by incorporating structural features in the document. The ELMo and BERT embeddings are also compared in both baseline and hierarchical structures. The results indicate that BERT embeddings generate a better document representation than ELMo in terms of model accuracy in this task. Meanwhile, the DAN models outperform others in generating document representation. In conclusion, the combination of hierarchical frameworks and contextual embeddings could significantly improve model performance in document classification.

There are potential improvements for each specific neural network structure. For instance, the RNN and CNN models can be implemented with attention mechanism. The DAN models could capture hierarchical information by increasing more FC layers. In the future work, this study could extend to different encoder structures in the hierarchical framework.

References

Jader Abreu, Luis Fred, David Macêdo, and Cleber Zanchettin. 2019. Hierarchical attentional hybrid neural networks for document classification. *arXiv preprint arXiv:1901.06610*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *European Chapter of the Association for Computational Linguistics EACL'17*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*.

Blaz Fortuna, Carolina Galleguillos, and Nello Cristianini. 2009. Detection of bias in media outlets with statistical learning methods. In *Text Mining*, pages 57–80. Chapman and Hall/CRC.

Shang Gao, Arvind Ramanathan, and Georgia Tourassi. 2018. Hierarchical convolutional attention networks for text classification. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).

Matthew Gentzkow. 2016. Polarization in 2016. *Toulouse Network for Information Technology Whitepaper*.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1113–1122.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 2019. Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. 2009. Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):721–735.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*.

Chenghua Lin, Yulan He, and Richard Everson. 2011. Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1153–1161.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Won, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. *Ijcai*.

Gregory J Martin and Ali Yurukoglu. 2017. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–99.

Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Matiur Rahman Minar and Jibon Naher. 2018. Violence originated from Facebook: A case study in Bangladesh. *arXiv preprint arXiv:1804.11241*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems.*

Emmanuel Vincent and Maria Mestre. 2018. Crowdsourced measure of news articles bias: Assessing contributors' reliability. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) co-located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zürich, Switzerland, July 5, 2018.*, pages 1–10.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 384–388.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430.*

Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):13.

Han Xiao. 2018. bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Wenpeng Yin and Hinrich Schütze. 2016. Multichannel variable-size convolution for sentence classification. *arXiv preprint arXiv:1603.04513.*

Ye Zhang, Stephen Roller, and Byron Wallace. 2016. Mgnc-cnn: A simple approach to exploiting multiple word embeddings for sentence classification. *arXiv preprint arXiv:1603.00968.*

Jianming Zheng, Fei Cai, Wanyu Chen, Chong Feng, and Honghui Chen. 2019. Hierarchical neural representation for document classification. *Cognitive Computation*, 11(2):317–327.

Bibliography

M. Abdul-Mageed and M. T. Diab. Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, volume 515, pages 3907–3914, 2012.

J. Abreu, L. Fred, D. Macêdo, and C. Zanchettin. Hierarchical attentional hybrid neural networks for document classification. *arXiv preprint arXiv:1901.06610*, 2019.

A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.

A. Agarwal, V. Sharma, G. Sikka, and R. Dhir. Opinion mining of news headlines using sentiwordnet. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, pages 1–5. IEEE, 2016.

A. Agarwal, R. Singh, and D. Toshniwal. Geospatial sentiment analysis using twitter data for uk-eu referendum. *Journal of Information and Optimization Sciences*, 39(1):303–317, 2018.

A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

M. T. AL-Sharuee, F. Liu, and M. Pratama. An automatic contextual analysis and clustering classifiers ensemble approach to sentiment analysis. *arXiv preprint arXiv:1705.10130*, 2017.

A. Aldayel and W. Magdy. Assessing sentiment of the expressed stance on social media. In *International Conference on Social Informatics*, pages 277–286. Springer, 2019.

S. Allan, B. Adam, and C. Carter. *Environmental risks and the media.* Psychology Press, 2000.

D. M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *technometrics*, 16(1):125–127, 1974.

F. Alqasemi, A. Abdelwahab, and H. Abdelkader. Opinion lexicon automatic construction on arabic language. 2018.

E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

A. Anderson. Media, politics and climate change: Towards a new research agenda. *Sociology compass*, 3(2):166–182, 2009.

L. Anthony. The cambridge dictionary of statistics. *Reference Reviews*, 2003.

L. Anthony, M. Edward, R. Seth, and K. John. Politics global warming. *New Haven*, 2018.

R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer, 2010.

S. E. Asch. Opinions and social pressure. *Scientific American*, 193(5):31–35, 1955.

A. Ashiquzzaman, A. K. Tushar, M. R. Islam, D. Shon, K. Im, J.-H. Park, D.-S. Lim, and J. Kim. Reduction of overfitting in diabetes prediction using deep learning neural network. In *IT Convergence and Security 2017*, pages 35–43. Springer, 2018.

A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)*, volume 1, pages 2–1. Citeseer, 2005.

S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.

A. Balahur, R. Steinberger, E. Van Der Goot, B. Pouliquen, and M. Kabadjov. Opinion mining on newspaper quotations. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 523–526. IEEE, 2009.

A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva. Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*, 2013.

C. Banea, R. Mihalcea, and J. Wiebe. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *LREC*, volume 8, pages 2–764, 2008.

P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

R. Barkemeyer, F. Figge, A. Hoepner, D. Holt, J. M. Kraak, and P.-S. Yu. Media coverage of climate change: An international comparison. *Environment and Planning C: Politics and Space*, 35(6):1029–1054, 2017.

M. Barnidge, A. C. Gunther, J. Kim, Y. Hong, M. Perryman, S. K. Tay, and S. Knisely. Politically motivated selective exposure and perceived media bias. *Communication Research*, 47(1):82–103, 2020.

R. A. Bauder and T. M. Khoshgoftaar. The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. *Health information science and systems*, 6(1):9, 2018.

R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin. An empirical study on class rarity in big data. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 785–790. IEEE, 2018.

F. R. Baumgartner, S. L. De Boef, and A. E. Boydstun. *The decline of the death penalty and the discovery of innocence*. Cambridge University Press, 2008.

P. Beattie and J. Milojevich. A test of the "news diversity" standard: Single frames, multiple frames, and values regarding the Ukraine conflict. *The International Journal of Press/Politics*, 22(1):3–22, 2017.

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2): 157–166, 1994.

W. L. Bennett. An introduction to journalism norms and representations of politics. 1996.

W. L. Bennett, R. G. Lawrence, and S. Livingston. *When the press fails: Political power and the news media from Iraq to Katrina*. University of Chicago Press, 2008.

D. Biber and E. Finegan. Adverbial stance types in english. *Discourse processes*, 11(1):1–34, 1988.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, 2007.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

T. Bolsen, J. N. Druckman, and F. L. Cook. How frames can undermine support for scientific adaptations: Politicization and the status-quo bias. *Public Opinion Quarterly*, 78(1):1–26, 2014.

A. E. Boydstun, D. Card, J. Gross, P. Resnick, and N. A. Smith. Tracking the development of media frames within and across policy issues. 2014.

M. Boykoff and G. Luedecke. Elite news coverage of climate change. In *Oxford Research Encyclopedia of Climate Science*. 2016.

M. T. Boykoff. The cultural politics of climate change discourse in UK tabloids. *Political Geography*, 27(5):549–569, 2008.

M. T. Boykoff. *Who speaks for the climate?: Making sense of media reporting on climate change*. Cambridge University Press, 2011.

T. Brants, F. Chen, and I. Tsochantaridis. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 211–218, 2002.

G. Brown and L. I. Kuncheva. "good" and "bad" diversity in majority vote ensembles. In *International workshop on multiple classifier systems*, pages 124–133. Springer, 2010.

C. Budak, S. Goel, and J. M. Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1): 250–271, 2016.

B. Burscher, R. Vliegenthart, and C. H. d. Vreese. Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, 34(5):530–545, 2016.

P. H. Calais Guerra, A. Veloso, W. Meira Jr, and V. Almeida. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM, 2011.

E. Cambria and A. Hussain. *Sentic computing: Techniques, tools, and applications*, volume 2. Springer Science & Business Media, 2012.

D. Card, A. Boydstun, J. H. Gross, P. Resnik, and N. A. Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, 2015.

A. Carvalho and J. Burgess. Cultural circuits of climate change in uk broadsheet newspapers, 1985–2003. *Risk Analysis: An International Journal*, 25 (6):1457–1469, 2005.

A. Caspi, K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. Harrington, J. McClay, J. Mill, J. Martin, A. Braithwaite, et al. Influence of life stress on depression: moderation by a polymorphism in the 5-htt gene. *Science*, 301(5631):386–389, 2003.

J.-R. Chang and Y.-S. Chen. Batch-normalized maxout network in network. *arXiv preprint arXiv:1511.02583*, 2015.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

J. Choo and S. Liu. Visual analytics for explainable deep learning. *IEEE computer graphics and applications*, 38(4):84–92, 2018.

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

C. M. Condit, A. Ferguson, R. Kassel, C. Thadhani, H. C. Gooding, and R. Parrott. An exploratory study of the impact of news headlines on genetic determinism. *Science Communication*, 22(4):379–395, 2001.

A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*, 2016.

A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E17-1104`.

J. Cook, N. Oreskes, P. T. Doran, W. R. Anderegg, B. Verheggen, E. W. Maibach, J. S. Carlton, S. Lewandowsky, A. G. Skuce, S. A. Green, et al. Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, 11(4):048002, 2016.

J. B. Corbett and J. L. Durfee. Testing public (un) certainty of science: Media representations of global warming. *Science Communication*, 26(2):129–151, 2004.

P. D'Angelo and J. A. Kuypers. *Doing news framing analysis: Empirical and theoretical perspectives.* Routledge, 2010.

D. K. Daniel. Market-driven journalism: Let the citizen beware? *Newspaper Research Journal*, 16(3):131, 1995.

K. Darwish, W. Magdy, and T. Zanouda. Improved stance prediction in a user similarity feature space. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 145–148, 2017.

K. Darwish, P. Stefanov, M. Aupetit, and P. Nakov. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152, 2020.

R. Das, M. Zaheer, and C. Dyer. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804, 2015.

J. J. Davis. The effects of message framing on response to environmental communications. *Journalism & Mass Communication Quarterly*, 72(2):285–299, 1995.

G. D. de Arruda, N. T. Roman, and A. M. Monteiro. An annotated corpus for sentiment analysis in political news. In *Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 101–110. SBC, 2015.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018a.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 2018b.

T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

P. DiMaggio, M. Nag, and D. Blei. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606, 2013.

X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240, 2008.

J. DiPeso. Media coverage and the environment: Why isn't global warming hot news? *Environmental Quality Management*, 16(1):97–103, 2006.

C. Dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, 2014.

J. N. Druckman, E. Peterson, and R. Slothuus. How elite partisan polarization affects public opinion formation. *American Political Science Review*, 107(1): 57–79, 2013.

E. Dumas-Mallet, A. Smith, T. Boraud, and F. Gonon. Poor replication validity of biomedical association studies reported by newspapers. *PloS one*, 12(2), 2017.

J. Durfee and J. Corbett. Context and controversy: Global warming coverage. *Nieman Reports*, 59(4):88, 2005.

P. S. Earle, D. C. Bowden, and M. Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), 2012.

J. Ebrahimi, D. Dou, and D. Lowd. A joint sentiment-target-stance model for stance classification in tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2656–2665, 2016.

H. Elfardy and M. Diab. Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 434–439, 2016.

J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

R. M. Entman. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173, 2007.

S. Feuerriegel, A. Ratku, and D. Neumann. Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 1072–1081. IEEE, 2016.

B. Fortuna, C. Galleguillos, and N. Cristianini. Detection of bias in media outlets with statistical learning methods. In *Text Mining*, pages 57–80. Chapman and Hall/CRC, 2009.

A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, pages 105–114. ACM, 2019.

S. M. Friedman, S. Dunwoody, and C. L. Rogers. Scientists and journalists. *AAAS, USA*, 1986.

W. A. Gamson, W. A. G. Gamson, W. A. Gamson, and W. A. Gamson. *Talking politics*. Cambridge university press, 1992.

S. Gao, A. Ramanathan, and G. Tourassi. Hierarchical convolutional attention networks for text classification. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2018.

R. K. Garrett. Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Journal of communication*, 59(4):676–699, 2009.

M. Gentzkow. Polarization in 2016. *Toulouse Network for Information Technology Whitepaper*, 2016.

F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999.

A. Giachanou and F. Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):1–41, 2016a.

A. Giachanou and F. Crestani. Opinion retrieval in twitter: is proximity effective? In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1146–1151, 2016b.

T. Gitlin. *The whole world is watching: Mass media in the making and unmaking of the new left.* Univ of California Press, 2003.

M. H. Goldberg, S. van der Linden, A. Leiserowitz, and E. Maibach. Perceived social consensus can reduce ideological biases on climate change. *Environment and Behavior*, page 0013916519853302, 2019.

Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, 2001.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

D. Gross and K. J. Miller. Adjectives in wordnet. *International Journal of lexicography*, 3(4):265–277, 1990.

A. Hamouda and M. Rohaim. Reviews classification using sentiwordnet lexicon. In *World congress on computer science and information technology*, volume 23, pages 104–105. sn, 2011.

J. Harrison. News. In F. (ed.), editor, *Pulling Newspapers Apart.* London, 2008.

J. Harrison. *The civil power of the news.* London: Palgrave., 2019.

S. Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91(1), 1991.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

J. Howard and S. Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

F. Huang and A. Yates. Distributional representations for handling sparsity in supervised sequence-labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 495–503. Association for Computational Linguistics, 2009.

A. Huertas and R. Kriegsman. Science or spin? *A report by the Union of concerned scientists, Washington, DC*, 12, 2014.

Y. Igarashi, H. Komatsu, S. Kobayashi, N. Okazaki, and K. Inui. Tohoku at semeval-2016 task 6: Feature-based model versus convolutional neural network for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 401–407, 2016.

T. L. Im, P. W. San, C. K. On, R. Alfred, and P. Anthony. Analysing market sentiment in financial news using lexical approach. In *2013 IEEE Conference on Open Systems (ICOS)*, pages 145–149. IEEE, 2013.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, 2014.

M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics* *and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691, 2015.

K. H. Jamieson and J. N. Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment.* Oxford University Press, 2008.

K. H. Jamieson, B. W. Hardy, and D. Romer. The effectiveness of the press in serving the needs of american democracy. 2007.

S. C. Jansen. Discovering the news: A social history of american newspapers, 1981.

A. E. Jasperson, D. V. Shah, M. Watts, R. J. Faber, and D. P. Fan. Framing and the public agenda: Media effects on the importance of the federal budget deficit. *Political Communication*, 15(2):205–224, 1998.

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.

Y. Jiang, X. Song, J. Harrison, S. Quegan, and D. Maynard. Comparing attitudes to climate change in the media using sentiment analysis based on latent dirichlet allocation. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 25–30, 2017.

Y. Jiang, J. Petrak, X. Song, K. Bontcheva, and D. Maynard. Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844, 2019a.

Y. Jiang, J. Petrak, X. Song, K. Bontcheva, and D. Maynard. Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence Representation Convolutional Network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844, 2019b.

Y. Jiang, Y. Wang, and X. S. D. Maynard. Comparing topic-aware neural networks for bias detection of news. In *Proceedings of 24th European Conference on Artificial Intelligence (ECAI 2020).* International Joint Conferences on Artificial Intelligence (IJCAI), 2020.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

N. Kalchbrenner, E. Grefenstette, and P. Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 355–363, 2006.

A. Kennedy and D. Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125, 2006.

J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. Potthast. SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval 2019).* Association for Computational Linguistics, 2019.

D. Kim, D. Seo, S. Cho, and P. Kang. Multi-co-training for document classification using various document representations: Tf–idf, lda, and doc2vec. *Information Sciences*, 477:15–29, 2019.

W. Y. Kim, J. S. Ryu, K. I. Kim, and U. M. Kim. A method for opinion mining of product reviews using association rules. In *Proceedings of the 2nd international conference on interaction sciences: Information technology, culture and human*, pages 270–274, 2009.

Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Knobloch-Westerwick and J. Meng. Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research*, 36(3):426–448, 2009.

P. Koehn. *Neural machine translation*. Cambridge University Press, 2020.

E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the omg! In *Fifth International AAAI conference on weblogs and social media*, 2011.

P. Krejzl and J. Steinberger. Uwb at semeval-2016 task 6: stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 408–412, 2016.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

J. A. Kuypers. *Presidential crisis rhetoric and the press in the post-Cold War world*. Greenwood Publishing Group, 1997.

M. Lan, C. L. Tan, J. Su, and Y. Lu. Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):721–735, 2009.

K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio. Batch normalized recurrent neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2657–2661. IEEE, 2016.

Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

H. W. Lee. Using twitter hashtags to gauge real-time changes in public opinion: an examination of the 2016 us presidential election. In *International Conference on Social Informatics*, pages 168–175. Springer, 2018.

M. S. Levendusky. Why do partisan media polarize viewers? *American Journal of Political Science*, 57(3):611–623, 2013.

J. Lewis and S. Cushion. The thirst to be first: An analysis of breaking news stories and their impact on the quality of 24-hour news coverage in the uk. *Journalism Practice*, 3(3):304–318, 2009.

J. Li, M.-T. Luong, and D. Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*, 2015.

T. Li, V. Sindhwani, C. Ding, and Y. Zhang. Bridging domains with words: Opinion analysis with matrix tri-factorizations. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 293–302. SIAM, 2010.

Z. Li, W. Shang, and M. Yan. News text classification model based on topic model. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–5. IEEE, 2016.

C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384, 2009.

C. Lin, Y. He, R. Everson, and S. Ruger. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data engineering*, 24(6):1134–1145, 2011.

H.-H. Lin and S.-F. Yang. An eye movement study of attribute framing in online shopping. *Journal of Marketing Analytics*, 2(2):72–80, 2014.

Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

S. Liu, L. Guo, K. Mays, M. Betke, and D. T. Wijaya. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, 2019a.

Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical word embeddings. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

C. Llewellyn, C. Grover, and J. Oberlander. Summarizing newspaper comments. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Won, and M. Cha. Detecting rumors from microblogs with recurrent neural networks. *Ijcai*, 2016.

D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, et al. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118, 2018.

D. Maynard and M. A. Greenwood. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC 2014 Proceedings*. ELRA, 2014.

D. Maynard, G. Gossen, A. Funk, and M. Fisichella. Should i care about your opinion? detection of opinion interestingness and dynamics in social media. *Future Internet*, 6(3):457–481, 2014.

K. McComas and J. Shanahan. Telling stories about global climate change: Measuring the impact of narratives on issue cycles. *Communication research*, 26(1):30–57, 1999.

M. McCombs. *Setting the agenda: Mass media and public opinion*. John Wiley & Sons, 2018.

M. McGlohon, N. Glance, and Z. Reiter. Star quality: Aggregating reviews to rank products and merchants. In *Fourth international AAAI conference on weblogs and social media*, 2010.

B. E. Meyerowitz and S. Chaiken. The effect of message framing on breast self-examination attitudes, intentions, and behavior. *Journal of personality and social psychology*, 52(3):500, 1987.

Z. Miao, Y. Li, X. Wang, and W.-C. Tan. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of The Web Conference 2020*, pages 617–628, 2020.

R. Mihalcea and H. Liu. A corpus-based approach to finding happiness. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 139–144, 2006.

R. Mihalcea and C. Strapparava. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 531–538. Association for Computational Linguistics, 2005.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

M. Mildenberger and D. Tingley. Beliefs about climate beliefs: the importance of second-order opinions for climate politics. *British Journal of Political Science*, 49(4):1279–1307, 2019.

M. R. Minar and J. Naher. Violence originated from Facebook: A case study in Bangladesh. *arXiv preprint arXiv:1804.11241*, 2018.

S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, 2016.

S. M. Mohammad, P. Sobhani, and S. Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23, 2017.

N. Naderi and G. Hirst. Classifying frames at the sentence level in news articles. *Policy*, 9:4–233, 2017.

J. Nagi, F. Ducatelle, G. A. Di Caro, D. Cireşan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 342–347. IEEE, 2011.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

S. Narayan, S. B. Cohen, and M. Lapata. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. *arXiv preprint arXiv:1808.08745*, 2018.

D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers. Analyzing entities and topics in news articles using statistical topic models. In *International conference on intelligence and security informatics*, pages 93–104. Springer, 2006.

D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

V.-A. Nguyen, J. Boyd-Graber, P. Resnik, and K. Miler. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1438–1448, 2015.

NSB. Science engineering indicators 2016. 2016. URL `https://www.nsf.gov/nsb/publications/2016/nsb20161.pdf`.

B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth international AAAI conference on weblogs and social media*, 2010.

A. E. Opperhuizen, K. Schouten, and E. H. Klijn. Framing a conflict! how media report on earthquake risks caused by gas drilling: A longitudinal analysis using machine learning techniques of media reporting on gas drilling from 1990 to 2015. *Journalism Studies*, pages 1–21, 2018.

J. Ortigosa-Hernández, J. D. Rodríguez, L. Alzate, M. Lucania, I. Inza, and J. A. Lozano. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing*, 92:98–115, 2012.

C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The measurement of meaning*. Number 47. University of Illinois press, 1957.

L. A. Overbey, S. C. Batson, J. Lyle, C. Williams, R. Regal, and L. Williams. Linking twitter sentiment and event data to monitor public opinion of geopolitical developments and trends. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 223–229. Springer, 2017.

A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

A. Panichella, B. Dit, R. Oliveto, M. Di Penta, D. Poshynanyk, and A. De Lucia. How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms. In *2013 35th International Conference on Software Engineering (ICSE)*, pages 522–531. IEEE, 2013.

S. Park, M. Ko, J. Kim, Y. Liu, and J. Song. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 113–122, 2011.

T. E. Patterson and W. Donsbagh. News decisions: Journalists as partisan actors. *Political communication*, 13(4):455–468, 1996.

W. Peng and D. H. Park. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

D. Peramunetilleke and R. K. Wong. Currency exchange rate forecasting from news headlines. In *Australian Computer Science Communications*, volume 24, pages 131–139. Australian Computer Society, Inc., 2002.

L. Perez and J. Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

D. Pérez-Granados, C. Lozano-Garzón, A. López-Urueña, and C. Jiménez-Guarín. Sentiment analysis in colombian online newspaper comments. In *Recent progress in data engineering and internet technology*, pages 113–119. Springer, 2012.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

D. T. Pham, S. S. Dimov, and C. D. Nguyen. Selection of k in k-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1):103–119, 2005.

M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

M. Potthast, T. Gollub, M. Wiegmann, and B. Stein. TIRA Integrated Research Architecture. In N. Ferro and C. Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer, 2019.

L. Y. Pratt. Discriminability-based transfer between neural networks. In *Advances in neural information processing systems*, pages 204–211, 1993.

V. Price, D. Tewksbury, and E. Powers. Switching trains of thought: The impact of news frames on readers' cognitive responses. *Communication research*, 24(5):481–506, 1997.

X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*, 2020.

M. Rafi, M. S. Shaikh, and A. Farooq. Document clustering based on topic maps. *arXiv preprint arXiv:1112.6219*, 2011.

H. Rahab, A. Zitouni, and M. Djoudi. Siaac: sentiment polarity identification on arabic algerian newspaper comments. In *Proceedings of the Computational Methods in Systems and Software*, pages 139–149. Springer, 2017.

S. D. Reese. The framing project: A bridging model for media research revisited. *Journal of communication*, 57(1):148–154, 2007.

S. D. Reese, O. H. Gandy Jr, and A. E. Grant. Covering the crisis in somalia: Framing choices by the new york times and the manchester guardian. In *Framing Public Life*, pages 191–200. Routledge, 2001.

A. Reeves, M. McKee, and D. Stuckler. 'it's the sun wot won it': Evidence of media influence on political attitudes and voting from a uk quasi-natural experiment. *Social science research*, 56:44–57, 2016.

R. E. Rice, A. Gustafson, and Z. Hoffman. Frequent but accurate: A closer look at uncertainty and opinion divergence in climate change print news. *Environmental Communication*, 12(3):301–321, 2018.

N. Risch, R. Herrell, T. Lehner, K.-Y. Liang, L. Eaves, J. Hoh, A. Griem, M. Kovacs, J. Ott, and K. R. Merikangas. Interaction between the serotonin transporter gene (5-httlpr), stressful life events, and risk of depression: a meta-analysis. *Jama*, 301(23):2462–2471, 2009.

V. Rubin, N. Conroy, Y. Chen, and S. Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17, 2016.

N. Ruchansky, S. Seo, and Y. Liu. Csi: A hybrid deep model for fake news detection. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM 17*, 2017. doi: 10.1145/3132847.3132877.

M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866, 2014.

A. Sadilek, H. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

H. Saif, Y. He, and H. Alani. Semantic sentiment analysis of twitter. In *International semantic web conference*, pages 508–524. Springer, 2012.

M. d. P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. A. Rodriguez-Garcia, and R. Valencia-Garcia. Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine*, 2017, 2017.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

C. N. d. Santos and V. Guimaraes. Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*, 2015.

H. A. Semetko and P. M. Valkenburg. Framing european politics: A content analysis of press and television news. *Journal of communication*, 50(2): 93–109, 2000.

A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

N. Shelke, S. Deshpande, and V. Thakare. Domain independent approach for aspect oriented sentiment analysis for product reviews. In *Proceedings of the 5th international conference on frontiers in intelligent computing: Theory and applications*, pages 651–659. Springer, 2017.

N. M. Shelke, S. Deshpande, and V. Thakre. Survey of techniques for opinion mining. *International Journal of Computer Applications*, 57(13):0975–8887, 2012.

M. Sherif. The psychology of social norms. 1936.

A. Singh, R. Nowak, and J. Zhu. Unlabeled data: Now it helps, now it doesn't. In *Advances in neural information processing systems*, pages 1513–1520, 2009.

V. K. Singh, R. Piryani, A. Uddin, and P. Waila. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *2013 International Mutli-Conference on Automation, Computing,* *Communication, Control and Compressed Sensing (iMac4s)*, pages 712–717. IEEE, 2013.

R. Slothuus and C. H. De Vreese. Political parties, motivated reasoning, and issue framing effects. *The Journal of Politics*, 72(3):630–645, 2010.

K. S. Smith, R. McCreadie, C. Macdonald, and I. Ounis. Analyzing disproportionate reaction via comparative multilingual targeted sentiment in twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 317–320, 2017.

P. Sollich and A. Krogh. Learning with ensembles: How overfitting can be useful. In *Advances in neural information processing systems*, pages 190–196, 1996.

R. Somaiya. How facebook is changing the way its users consume journalism. *The New York Times*, 26:14, 2014.

X. Song, J. Petrak, Y. Jiang, I. Singh, D. Maynard, and K. Bontcheva. Classification aware neural topic model and its application on a new covid-19 disinformation corpus. *arXiv preprint arXiv:2006.03354*, 2020.

A. Spence and N. Pidgeon. Framing and communicating climate change: The effects of distance and outcome frame manipulations. *Global Environmental Change*, 20(4):656–667, 2010.

K. R. Stamm, F. Clark, and P. R. Eblacas. Mass communication and public understanding of environmental problems: the case of global warming. *Public understanding of science*, 9(3):219–238, 2000.

J. Steinberger, T. Brychcín, and M. Konkol. Aspect-level sentiment analysis in czech. In *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 24–30, 2014.

V. Štětka. Media ownership and commercial pressures. *Media and Democracy in Central and Eastern Europe*, 2013.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2): 111–133, 1974.

N. J. Stroud. Polarization and partisan selective exposure. *Journal of communication*, 60(3):556–576, 2010.

C. Sun, X. Qiu, Y. Xu, and X. Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer, 2019.

X. Sun, X. Ren, S. Ma, and H. Wang. meprop: Sparsified back propagation for accelerated deep learning with reduced overfitting. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3299–3308. JMLR. org, 2017.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

J. W. Tankard Jr. The empirical approach to the study of media framing. In *Framing public life*, pages 111–121. Routledge, 2001.

N. Terkildsen and F. Schnell. How media frames move public opinion: An analysis of the women's movement. *Political research quarterly*, 50(4):879–900, 1997.

A. Trabelsi and O. Zaiane. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

C. Trumbo. Constructing climate change: claims and frames in us news coverage of an environmental issue. *Public understanding of science*, 5(3):269–284, 1996.

B. Tsolmon, A.-R. Kwon, and K.-S. Lee. Extracting social events based on timeline and sentiment analysis in twitter corpus. In *International Conference on Application of Natural Language to Information Systems*, pages 265–270. Springer, 2012.

O. Tsur, D. Calacci, and D. Lazer. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, 2015.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 2010.

P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

S. Unankard, X. Li, M. Sharaf, J. Zhong, and X. Li. Predicting elections from social networks based on sub-event detection and sentiment analysis. In *International Conference on Web Information Systems Engineering*, pages 1–16. Springer, 2014.

S. Ungar. The rise and (relative) decline of global warming as a social problem. *Sociological quarterly*, 33(4):483–501, 1992.

S. Ungar. Knowledge, ignorance and the popular culture: climate change versus the ozone hole. *Public Understanding of Science*, 9(3):297–312, 2000.

M. Unnisa, A. Ameen, and S. Raziuddin. Opinion mining on twitter data using unsupervised learning technique. *International Journal of Computer Applications*, 148(12):975–8887, 2016.

J. E. Van Engelen and H. H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

A. Vasalou, A. N. Joinson, and D. Courvoisier. Cultural differences, experience with social networks and the nature of “true commitment” in facebook. *International journal of human-computer studies*, 68(10):719–728, 2010.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017a.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems.*, 2017b.

J. J. Väyrynen, T. Honkela, and L. Lindqvist. Towards explicit semantic features using independent component analysis. In *Unknown host publication*, pages 20–27. 2007.

E. Vincent and M. Mestre. Crowdsourced measure of news articles bias: Assessing contributors' reliability. In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) co-located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zürich, Switzerland, July 5, 2018.*, pages 1–10, 2018a. URL `http://ceur-ws.org/Vol-2276/paper1.pdf`.

E. Vincent and M. Mestre. Crowdsourced measure of news articles bias: Assessing contributors' reliability. In *SAD/CrowdBias@ HCOMP*, pages 1–10, 2018b.

H. T. Vu, Y. Liu, and D. V. Tran. Nationalizing a global phenomenon: A study of how the press in 45 countries and territories portrays climate change. *Global Environmental Change*, 58:101942, 2019.

G. Wang, J. Hao, J. Ma, and H. Jiang. A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, 38(1):223–230, 2011.

H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations*, pages 115–120. Association for Computational Linguistics, 2012.

J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019.

Y. Wang and W. Xu. Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, 105:87–95, 2018.

S. R. Weart. *The discovery of global warming*. Harvard University Press, 2008.

W. Wei, X. Zhang, X. Liu, W. Chen, and T. Wang. pkudblab at semeval-2016 task 6 : A specific convolutional neural network system for effective stance detection. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016. doi: 10.18653/v1/s16-1062.

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

G. Wiedemann, S. Remus, A. Chawla, and C. Biemann. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*, 2019.

L. Wilkins. Between facts and values: Print media coverage of the greenhouse effect, 1987-1990. *Public understanding of science*, 2(1):71–84, 1993.

K. M. Wilson. Communicating climate change through the media. *Environmental risks and the media*, pages 201–217, 2000.

M.-S. Won and J.-H. Lee. Embedding for out of vocabulary words considering contextual and morphosyntactic information. In *2018 International Conference on Fuzzy Theory and Its Applications (iFUZZY)*, pages 212–215. IEEE, 2018.

H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):13, 2008.

M. Wu, F. Liu, and T. Cohn. Evaluating the utility of hand-crafted features in sequence labelling. *arXiv preprint arXiv:1808.09075*, 2018.

X. Wu, L. Fang, P. Wang, and N. Yu. Performance of using LDA for Chinese news text classification. In *2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1260–1264. IEEE, 2015.

H. Xiao. bert-as-service. `https://github.com/hanxiao/bert-as-service`, 2018.

H. Xu, M. Dong, D. Zhu, A. Kotov, A. I. Carcone, and S. Naar-King. Text classification with topic-based word embedding and convolutional neural networks. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 88–97. ACM, 2016.

Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*, 2016a.

Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016b.

L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946, 2009.

W. Yin and H. Schütze. Multichannel variable-size convolution for sentence classification. *arXiv preprint arXiv:1603.04513*, 2016.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

N. Yu and S. Kübler. Filling the gap: Semi-supervised learning for opinion detection across domains. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 200–209. Association for Computational Linguistics, 2011.

R. Zamith, J. Pinto, and M. E. Villar. Constructing climate change in the Americas: An analysis of news coverage in US and South American newspapers. *Science Communication*, 35(3):334–357, 2013.

J. Zeng, W. K. Cheung, and J. Liu. Learning topic models by belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (5):1121–1134, 2012.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

Y. Zhang and B. Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.

Z. Zhang, Q. Ye, Z. Zhang, and Y. Li. Sentiment classification of internet restaurant reviews written in cantonese. *Expert Systems with Applications*, 38(6):7674–7682, 2011.

R. Zhao and K. Mao. Fuzzy bag-of-words model for document representation. *IEEE Transactions on Fuzzy Systems*, 26(2):794–804, 2018.

J. Zheng, F. Cai, W. Chen, C. Feng, and H. Chen. Hierarchical neural representation for document classification. *Cognitive Computation*, 11(2):317–327, 2019.

C. Zhou, C. Sun, Z. Liu, and F. Lau. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.

J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020.

D. Zillmann, L. Chen, S. Knobloch, and C. Callison. Effects of lead framing on selective exposure to internet news reports. *Communication research*, 31 (1):58–81, 2004.