University of Sheffield

Multilingual and Multi-Domain Rumour Stance Classification

Yue Li

Supervisors: Carolina Scarton, Zhixue Zhao, Kalina Bontcheva

A thesis submitted for the degree of Doctor of Philosophy in Computer Science

in the

School of Computer Science

February 25, 2026

Declaration

I, Yue Li, hereby declare that this thesis is my own original work and has not been submitted, in whole or in part, for any previous application for a degree or qualification. Except where states by acknowledgment or reference, the work in this thesis presented is entirely my own.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr Carolina Scarton, for her continuous encouragement, support, and guidance throughout both my Master's and PhD studies. She is always patient and professional, and it's been a great honor to work with her over years. Her professionality and encouragement motivate me to become a better researcher. Her understanding and support in my personal life have also meant a great deal to me. I deeply appreciate her not only as an outstanding supervisor but also as a valued friend. Words cannot fully express my gratitude to Carol who has supported and inspired me throughout this journey. I sincerely wish her continued success in her professional career and happiness in personal life.

I am also profoundly grateful to Dr Zhixue Zhao and Professor Kalina Bontcheva for their academic and emotional support during my doctoral journey. I have benefited greatly from Zhixue's invaluable and practical advice on research, paper writing, and rebuttals, which has significantly shaped my development as a researcher. I would like to also express my gratitude to my examiners, Professor Nikos Aletras and Arkaitz Zubiaga, who provided insightful suggestions for this thesis.

My sincere thanks also go to my colleagues in the GATE team: Dr Xingyi Song, Professor Diana Maynard, Ian Roberts, Mark Greenwood, Fatima Haouari, Olesya Razuvayevskaya, Iknoor Singh, Jake Vasilakes, Siqi Sun, Ben Wu, Mali Jin, Mugdha Pandya and Joao A Leite, for creating such a supportive and enjoyable working environment. I have greatly valued the collaborative atmosphere, team lunches, and the generous technical assistance and advice.

Finally, I would like to express my deepest love and gratitude to my family, my cousin and my best friends, Hongsong Wang, Zhixue Zhao, Siqi Sun, Xinyuan Kang, Chunyu Deng, Zheng Liu and Huiyin Xue, for their valuable support and encouragement. Your friendship and care have illuminated my PhD journey. Thank you for always being by my side and cheering me up, especially during my difficult moments, and for sharing countless happy time that I will treasure forever.

Abstract

Rumour stance classification focuses on a conversation initialised by a rumour-related source post on social media, aiming to determine the stance of each reply's author towards the rumour. Accurately capturing public stance facilitates assessing the verdict or check-worthiness of the information. Given the diversity and multilingual nature of online rumours, systems must be able to generalise and adapt to diverse domains and languages. This thesis addresses these issues by investigating methods for improving the generalisation and adaptation of rumour stance classification across new rumours, domains, and languages.

Firstly, we identify the distinction between rumour stance classification and generic stance classification by analysing the special role of the stance target (i.e, rumour) in model generalisation. We propose a new ensemble-based method to enhance the reasoning with rumours, achieving state-of-the-art performance. Secondly, we assess the generalisability of top-performing models under domain shift, and propose a LLM-assisted self-training framework for effective adaptation without access to both source and target domain labelled data. Thirdly, we reveal how class labels design in prompts affect LLMs' generalisation in zero-shot in-Context Learning (ICL) (e.g, the lexical choice between "agree" and "support" for positive stance), and introduce an efficient post-hoc method for optimal label selection.

Furthermore, this thesis investigates the adaptation of English-centric rumour stance classification models to non-English languages. We create the largest multilingual benchmark dataset with nine diverse high- and medium-resource languages. We then reveal performance inconsistency across languages and further analyse strategies to improve model performance.

Contents

List of Figures

List of Tables

Chapter 1

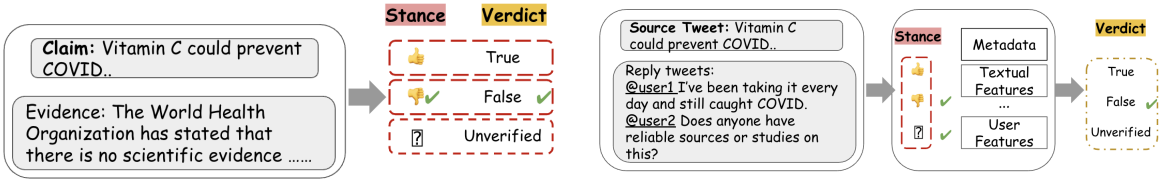
Introduction

Although social media platforms enable faster and broader information dissemination than traditional media (Phuvipadawat & Murata 2010, Lazer et al. 2009), it also results in a large number of unfiltered, potentially false, misleading or even malicious contents. Although efforts have been made by journalists, fact-checking organisations and social media platforms to verify and limit the spreading of such contents, verification is inherently time-sensitive, and manual approaches are insufficient to manage the vast amount of online information. Consequently, Natural Language Processing (NLP) tasks, frameworks and methods have been developed to assist in the verification and analysis of online mis- or disinformation[1], e.g., fact-checking (Thorne & Vlachos 2018, Augenstein et al. 2019, Mihaylova et al. 2018, Abu Ahmad et al. 2025, Liu, Das, Boltz, Zhou, Pinaroc, Lease & Lee 2024), rumour verification (Zubiaga et al. 2018*a*, Derczynski et al. 2017, Gorrell et al. 2019*a*), check-worthiness (Gencheva et al. 2017, Atanasova et al. 2018, Nakov, Da San Martino, Elsayed, Barrón-Cedeno, Míguez, Shaar, Alam, Haouari, Hasanain, Babulkov et al. 2021), and narrative analysis and extraction (Li et al. 2023, Nikolaidis et al. 2025, Glavaš et al. 2014, Haouari et al. 2025).

Stance classification plays an essential role in computational approaches for detecting and analysing online misinformation and rumours (Hardalov et al. 2022*b*), which has been broadly framed as either (1) a standalone fact-checking task (Thorne & Vlachos 2018, Hossain et al. 2020, Hanselowski et al. 2019, Chen et al. 2020); or (2) an intermediate step within multi-stage misinformation/rumour analysis and verification pipelines (Ferreira & Vlachos 2016, Zubiaga et al. 2018*a*, Derczynski et al. 2017, Gorrell et al. 2019*a*, Zheng et al. 2022), as illustrated in Figure 1.1. In the former formulation (Figure 1.1a), the stance of a single or multiple pieces of evidence (e.g., Wikipedia articles or web pages) towards the target claim is directly interpreted as the verdict of the claim (Thorne & Vlachos 2018, Augenstein et al. 2019). In the latter formulation (Figure 1.1b), the stances expressed in social media posts with respect to a claim or rumour are aggregated and combined with additional contextual signals such as user metadata and interaction networks, which are subsequently exploited by verification pipelines.

This thesis addresses (2), i.e., the task of **rumour stance classification** on social media (Zubiaga et al. 2018*a*). As shown in Figure 1.2, given an online conversation initialised by a

[1]Misinformation is accidentally false while disinformation refers to deliberately false information. Rumours can be either misinformation or disinformation, although their veracity is not necessary to be false (Zubiaga et al. 2018*a*). In this thesis, we focus on rumours without consideration of their underlying intent.



(a) *Stance classification as fact-checking.*

(b) *Stance classification in a verification pipeline.*

Figure 1.1: *Illustrations of the common two stance classification formulations in misinformation, disinformation or rumour analysis.*

rumour, rumour stance classification aims to identify the attitude or perspective of the reply's author towards the rumourous source post. Typically, the following four stance categories are considered:

- *Support*: the author of the reply agrees with the source post.
- *Deny*: the author of the reply disagrees with the source post.
- *Query*: the author of the reply asks for more evidence regarding the veracity of the source post.
- *Comment*: the author of the reply makes their own comment without contribution to assessing the veracity of the rumour.

Rumour stance classification has a wide range of practical applications. It is a key component of automatic rumour verification pipelines on social media (Zubiaga et al. 2018*a*), where stance information is either incorporated as features within rumour verification models or jointly modelled with the rumour verification task itself (Dungs et al. 2018, Dougrez-Lewis et al. 2021). Beyond integration into verification pipelines, rumour stance classification itself is also valuable as a stand-alone tool for journalists and fact-checkers[2]. It can help highlight contents that are controversial, disputed, or likely to be misleading, efficiently guiding journalists and fact-checkers towards check-worthy information. Furthermore, it could provide useful insights for researchers and stakeholders into public attitudes and discourse surrounding circulating rumours. For instance, it can facilitate studies on patterns of rumour propagation, detection of echo chambers or polarization, and analysis of how different communities respond to rumours. In crisis management, it can also assist authorities in tracking misinformation and coordinating timely interventions.

Despite its importance, rumour stance classification presents substantial challenges due to the variability of rumours across domains. Models trained on specific datasets often generalise poorly to new rumours in different events or topics (Kochkina et al. 2023), since domain shift emerges from linguistic, topical and knowledge divergences, with stance expressed significantly different across rumours, for instance, between health-related (e.g., COVID-19, vaccination) and political-related (e.g., elections) rumours. Language adaptation poses a further critical

[2]e.g., Our multilingual rumour stance classification model deployed at the Verification Plugin (`https://weverify.eu/verification-plugin/`).

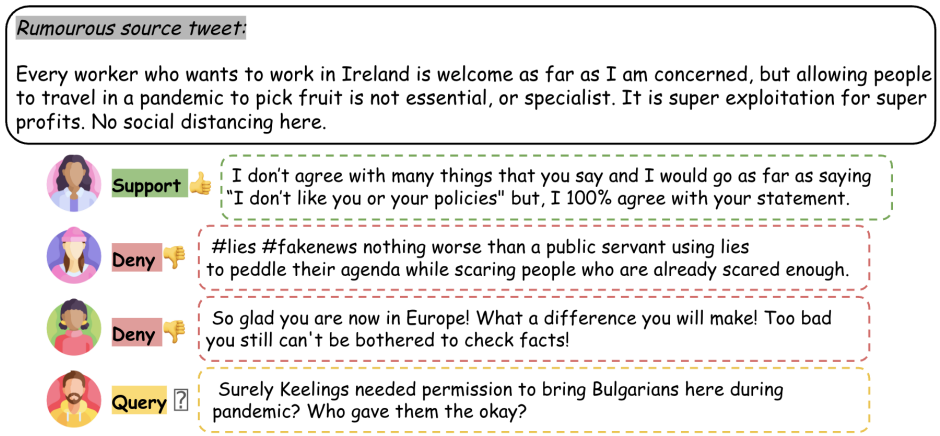


Figure 1.2: *Example of a Twitter conversation discussing a rumour spread in Ireland, where the label associated with each reply is the stance towards the source tweet.*

challenge for rumour stance classification, since misinformation, disinformation and rumours are global phenomena that spread across diverse linguistic communities.

Existing studies on rumour stance classification have primarily focused on improving generalisation of supervised models through feature engineering (Aker et al. 2017, Bahuleyan & Vechtomova 2017, Ghanem et al. 2019, García Lozano et al. 2017), effective modelling of the conversation structure (Kochkina et al. 2017, Kumar & Carley 2019, Li, Sujana & Kao 2020, Wei et al. 2019, Li, Sujana & Kao 2020, Yang et al. 2019, Fajcik et al. 2019) and imbalanced data treatment (Li & Scarton 2020, García Lozano et al. 2017, Ghanem et al. 2019, Yang et al. 2019, Singh et al. 2017). However, current approaches are exclusively evaluated on RumourEval and PHEME datasets (Derczynski et al. 2017, Gorrell et al. 2019*a*) that exhibit significant overlap. This leaves a significant gap in the understanding of performance degradation and adaptation under domain shift. Rumour stance classification is also incorporated into broader multi-task and multi-dataset stance classification research (Hardalov et al. 2021, Arakelyan et al. 2023), where the inherent differences between rumour stance classification and generic stance classification are often overlooked. Furthermore, existing research remains predominantly English-centric due to the limited availability of non-English datasets (Zubiaga et al. 2016, Lozhnikov et al. 2018, Lillie et al. 2019*a*, Zheng et al. 2022).

This thesis addresses the above research gaps and contributes to the research field through: (1) identifying the fundamental distinctions between rumour stance classification and generic stance classification, and proposing a novel perspective to enhance model generalisability; (2) introducing two new datasets that facilitate studies for domain and language adaptations; (3) proposing three advanced and data-efficient methods for model adaptation across domains and languages under supervision and ICL.

1.1 Research Aims and Objectives

This thesis focuses on the generalisation and adaptation of rumour stance classification across domains and languages. We aim to achieve the following research objectives:

- Intuitively, the *target* of the stance (i.e., the rumour in rumour stance classification) is expected to be an essential component in stance classification (Kaushal et al. 2021). However, rumour stance classification contains real-world data that could be naturally target-independent. For instance, a reply that directly responds to a rumour with *"This is fake news"* clearly conveys disagreement, regardless of the specific content of the rumour (also see the two *deny* examples in Figure 1.2). This characteristic makes rumour stance classification distinct from generic stance classification tasks which typically assume a strong dependency on the *target* when improving model generalisaion (Xu et al. 2018, Clark et al. 2021, He et al. 2022). Therefore, **we aim to uncover the true role of the *target* in rumour stance classification when generalising to new rumours, and develop new models that could flexibly reason with the *target*.**

- Studies in rumour stance classification have so far been shaped by three closely related datasets with notable overlapping contents and similar stance distributions: PHEME (Zubiaga et al. 2016), RumourEval 2017 (Derczynski et al. 2017) and RumourEval 2019 (Gorrell et al. 2019*a*). Specifically, the RumourEval 2017 dataset constitutes the English subset of the PHEME dataset, and RumourEval 2019 dataset reuses RumourEval 2017 as its Twitter training set. As a result, current top-performing models have not been tested for real-world generalisation, where domain and label distributions could differ substantially. Adaptation to such shifts also remains underexplored. Therefore, **we aim to develop a new dataset suitable to evaluate model generalisation and propose new methods that could facilitate data-efficient domain adaptation.**

- Different stance classification datasets often adopt varying label inventories (e.g., *agree-disagree* vs. *favor-against* for positive and negative stances), a challenge addressed as label adaptation in prior work on cross-dataset stance classification (Hardalov et al. 2021). However, this variability also leads to arbitrary lexical choices for stance label names in zero-shot ICL, where model generalisation is sensitive to the wording of the labels (Mu et al. 2024). Therefore, **we aim to develop methods that can effectively and efficiently identify optimal label sets to improve model generalisation to rumour stance classification in zero-shot ICL.**

- For high- and medium-resource languages, transferring knowledge learnt from English rumour stance classification datasets to a target language by fine-tuning MLMs (e.g., multilingual BERT (Pires et al. 2019)) has been studied by prior work, mainly for three languages: Danish, Russian, and German (Hardalov et al. 2022*a*, Scarton & Li 2021). Nevertheless, recent advances in LLMs (Grattafiori et al. 2024, Le Scao et al. 2023) have offered support for many other high- and medium-resource languages (e.g., German and Hindi). Therefore, **we aim to develop a large scale multilingual dataset, and reveal how LLMs' multilingual capability can be leveraged to non-English rumour stance classification.**

Overall, the aim of this thesis is to advance the generalisation capability of rumour stance classification systems across rumours, domains and languages. By addressing model generalisation and adaptation from multiple perspectives, the thesis aims to contribute towards more adaptable, scalable, efficient and linguistically inclusive rumour stance classification systems.

Specifically, we seek to address the following four primary research questions for generalisation and adaptation of rumour stance classification:

RQ1: What is the role of targets in the generalisation of rumour stance classification? How can models adequately use information from targets?

RQ2: Can state-of-the-art rumour stance classification models effectively generalise across domains? How can their performance be improved via adaptation to new domains?

RQ3: In ICL with LLMs, to what extent does the label design impact the model's ability to generalise to rumour stance classification? How can such generalisation be enhanced through label optimisation?

RQ4: How can English-centric models effectively adapt to non-English languages for rumour stance classification?

1.2 Thesis Overview: Publications and Contributions

This section lists the contributions of this thesis. It follows a *thesis by publications* format and comprises a collection of five papers, with each paper corresponding to a separate chapter.

Publication I: Can We Identify Stance Without Target Arguments? A Study for Rumour Stance Classification In this publication, we address RQ1 by improving supervised model's generalisability to unseen rumours. We analyse the special role of the *target* in generalisation to unseen rumours for rumour stance classification, and consequently propose a novel method to enhance model's reasoning with the rumour. The contributions of this publication are:

- Supplementary annotations on the current largest English rumour stance classification dataset, where we manually categorise the replies into *target-dependent* (i.e. rumour is essential for stance inference) and *target-independent* (i.e. rumour is unnecessary for stance inference), facilitating fine-grained model evaluation and analysis.
- Empirical analysis that highlights the fundamental distinction between generic stance classification and rumour stance classification.
- A novel method that improves the reasoning with rumour, exhibiting better generalisation to unseen rumours for both target-dependent and -independent replies.

This work has been published in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (Li & Scarton 2024). My contributions to the work: conceptualization, data annotation, methodology, validation and writing.[3]

[3]Part of this work required ethical approval, which was granted by the University of Sheffield (application number 046997).

Publication II: Rumour Stance Classification Adaptation Under Domain Shift: a case study of rumours in Ireland In this publication, we address RQ2 by first assessing the generalisability of current rumour stance classification models, and then propose a data-efficient adaptation framework. The contributions of this publication are:

- The first English rumour stance classification benchmarking test set that significantly differs from the RumourEval datasets in topic, vocabulary and stance distribution, making it well-suited for evaluation and adaptation under domain shift.
- The first generalisability analysis of rumour stance classification models.
- A novel LLM-assisted self-training framework enabling effective and efficient adaptation without access to labelled data in both source and target domains.

This work is currently under review in the EPJ Data Science Journal. My contributions to the work: conceptualization, data collection and annotation, methodology, validation and writing.[4]

Publication III: Label Set Optimization via Activation Distribution Kurtosis for Zero-Shot Classification with Generative Models In this publication, we address RQ3 by improving LLMs' generalisation to rumour stance classification in ICL. We propose a post-hoc method for selecting optimal label sets (e.g., lexical choice between *agree* and *support* for rumour stance classification) in zero-shot ICL with LLMs, built upon the observations in our empirical analysis. The contributions of this publication are:

- The first benchmark on how variants of label options (i.e., lexical choice, order, and elaboration) in prompts affect zero-shot ICL models' performance for classification tasks.
- Empirical demonstration that zero-shot ICL performance negatively correlates with the number of outlier neurons in feed-forward neuron networks when varying the lexical choices for label options.
- A novel and efficient post-hoc method for optimal label selection in zero-shot ICL.

This work has been accepted to appear at The 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)[5] (Li et al. 2024) My contributions to the work: conceptualization, methodology, validation and writing.[6]

Publication IV: SCRum-9: Multilingual Stance Classification over Rumours on Social Media In this publication, we address RQ4 with the focus on high- and medium-resource non-English languages. We introduce a new multilingual rumour stance classification dataset. We then benchmark and analyse language adaptation of multilingual LLMs and masked language models (MLMs) in both ICL and fine-tuning setups. We also examine the

[4]Part of this work required ethical approval, which was granted by the University of Sheffield (application number 046997).

[5]Top 15% of accepted papers.

[6]Part of this work required ethical approval, which was granted by the University of Sheffield (application number 059099).

relationship between model predictions and human uncertainty. The contributions in this publication are:

- The current largest multilingual rumour stance classification dataset covering Czech, German, English, Spanish, French, Hindi, Polish, Portuguese, and Russian. Given the subjectivity of rumour stance classification, the dataset was also annotated with an additional design to incorporate annotator uncertainty.
- Extensive analyses of strategies for improving ICL performance in non-English languages.
- The first work to examine the effectiveness of synthetic multilingual data generated by LLMs for rumour stance classification.
- Analysis of the relationship between model predictions and human uncertainty on ambiguous cases.

This work is accepted to appear at The 20th International AAAI Conference on Web and Social Media Understanding the World Through the Web (ICWSM 2026) (Li, Vasilakes, Zhao & Scarton 2025). My contributions to the work: conceptualization, data annotation, methodology, validation and writing.[7]

Conclusions We summarise our findings and contributions presented in previous chapters of this thesis, outline potential directions for future research.

1.3 Other Research Contributions

We list other research contributions that are not well aligned with the thesis topics as follows:

1. Benedetta Muscato, **Yue Li**, Gizem Gezici, Zhixue Zhao, and Fosca Giannotti. Seeing All Sides: Multi-Perspective In-Context Learning for Subjective NLP. To appear in EACL 2026.
2. **Yue Li**, Zhixue Zhao, and Carolina Scarton. 2025. It's All About In-Context Learning! Teaching Extremely Low-Resource Languages to LLMs. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29544–29559, Suzhou, China. Association for Computational Linguistics.
3. **Yue Li**, Carolina Scarton, Xingyi Song, and Kalina Bontcheva. 2023. Classifying COVID-19 Vaccine Narratives. In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pages 648–657, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
4. Ben Wu, **Yue Li**, Yida Mu, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Don't waste a single annotation: improving single-label classifiers through soft labels. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5347–5355, Singapore. Association for Computational Linguistics.

[7]Part of this work required ethical approval, which was granted by the University of Sheffield (application number 059099).

5. Iknoor Singh, **Yue Li**, Melissa Thong, and Carolina Scarton. 2022. GateNLP-UShef at SemEval-2022 Task 8: Entity-Enriched Siamese Transformer for Multilingual News Article Similarity. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1121–1128, Seattle, United States. Association for Computational Linguistics.

Chapter 2

Publication I: Can We Identify Stance Without Target Arguments? A Study for Rumour Stance Classification

Overview In this chapter, we study the generalisability of supervised rumour stance classification models to rumours that are not included during training. We analyse the special role of the target (i.e., rumour) in model generalisation. We empirically identify the presence of target-independent replies and demonstrate how they result in the limitation of the current target-centric modelling approaches that assume stance is always strictly conditioned on the specific rumour. To address this issue, we propose a novel ensemble-based framework that models the target in a more flexible manner, allowing the system to better handle both target-dependent and target-independent cases. Our approach achieves state-of-the-art performance on the RumourEval datasets.

Can We Identify Stance Without Target Arguments? A Study for Rumour Stance Classification

Yue Li, Carolina Scarton

School of Computer Science, University of Sheffield, UK

Abstract

Considering a conversation thread, rumour stance classification aims to identify the opinion (e.g. agree or disagree) of replies towards a *target* (rumour story). Although the target is expected to be an essential component in traditional stance classification, we show that rumour stance classification datasets contain a considerable amount of real-world data whose stance could be naturally inferred directly from the replies, contributing to the strong performance of the supervised models without awareness of the target. We find that current target-aware models underperform in cases where the context of the target is crucial. Finally, we propose a simple yet effective framework to enhance reasoning with the targets, achieving state-of-the-art performance on two benchmark datasets.

u1(Target): BREAKING NEWS: New York Times is reporting the Canadian soldier who was shot has died from their injuries. Heartbreaking.
u2(T-I) : @u1 Canadian news contradicts this.
u3(T-D): @u1 CTV just said he is being treated and is stable?

Figure 2.1: *Example of Target-Independent (T-I) and Target-Dependent (T-D) direct replies that deny a target from Gorrell et al. (2019*a*).*

2.1 Introduction

Automatic stance classification that aims to identify the type of an expressed opinion towards a single or multiple *targets*, plays a key role in many Natural Language Processing (NLP) applications, such as rumour analysis (Zubiaga et al. 2016). A target could be a person, an organisation, or rumour story, depending on the use case (Hossain et al. 2020, Zubiaga et al. 2016, Ferreira & Vlachos 2016, Allaway & McKeown 2020). The target plays a fundamental role in stance classification, being expected to appear either explicitly or implicitly, making it a key difference from sentiment analysis that can be framed as target-independent (Küçük & Can 2020, Liu, Lin, Ji, Li, Fu & Wang 2022).

Previous work shows that a BERT-based model, without awareness of the target, achieves comparable or even better performance than target-aware models on many stance classification datasets, due to spurious sentiment- and lexicon-stance correlations in the training sets (Kaushal et al. 2021). Similar results are observed in other context-dependent tasks, such as Natural Language Inference and Argument Reasoning Comprehension, where models without background knowledge achieve an impressive performance due to spurious or superficial cues in the datasets (Poliak et al. 2018, Niven & Kao 2019).

In this paper, we further analyse the above phenomenon for *rumour stance classification* on Twitter. Given a conversation initialised by a rumourous *source tweet*, this task aims to classify the stance of each reply towards the rumour into *support*, *deny*, *query* and *comment*.[1] The vagueness and lack of specificity in the reply tweets result in the disparity between rumour stance classification and traditional stance classification datasets. For instance, in Figure 2.1, one can reasonably deduce that the reply from *u2* disagrees with the target before reading the content of the target. This is in contrast to traditional stance classification where the stance may vary for different targets, making it always essential to consider them (e.g., Sobhani et al. 2017, Conforti et al. 2020).

We empirically show that the strong behaviour of models without awareness of the target (dubbed *target-oblivious*) could be explained by the existence of the reply posts whose stance can be naturally inferred without knowing the target.[2] More importantly, we demonstrate that current state-of-the-art target-aware models lack reasoning with the target, performing unexpectedly poorly on the cases when the target is necessary. Based on our observations, we

[1]The target of rumour stance classification is the rumour story by task definition, but these are not given in the datasets. Instead, they are implied by the source tweets. In this work we use the terminology *target* to indicate the source tweet, because it is treated as the *target* in data annotation and applications (Zubiaga et al. 2016, Hardalov et al. 2022*b*, Kaushal et al. 2021)

[2]Annotations can be found at: `https://github.com/YLi999/Target-Annotations-RumourEval`

propose a simple yet effective framework which would benefit from the target-oblivious model and would also enhance the reasoning with the targets.

2.2 The Role of Target Arguments

We conduct an annotation study by categorising the replies into *target-dependent* (i.e. target is essential for stance inference) and *target-independent* (i.e. target is unnecessary for stance inference). We then evaluate various models trained with or without awareness of the target (i.e. *target-aware* and *target-oblivious* models).

2.2.1 Data Annotation

Dataset Three established English datasets are available for rumour stance classification on social media: *PHEME* (Zubiaga et al. 2016), *RumourEval 2017* (Derczynski et al. 2017) and *RumourEval 2019* (Gorrell et al. 2019*a*). RumourEval 2017 consists of the English PHEME dataset, and RumourEval 2019 is an extension of the 2017 dataset. Therefore, we consider the largest RumourEval 2019 dataset.[3] We present the corresponding stance label distribution of RumourEval 2019 Twitter dataset in Table 2.1.

Dataset	Support	Deny	Query	Comment
Training	633 (15%)	335 (8%)	358 (8%)	2896 (69%)
Validation	68 (7%)	69 (7%)	106 (10%)	778 (76%)
Testing	91 (9%)	92 (9%)	56 (6%)	771 (76%)

Table 2.1: *Stance label distribution of RumourEval 2019 Twitter dataset.*

RumourEval 2019 training and validation sets consist of conversations regarding rumour stories which emerged during breaking news (e.g., Germanwings plane crash, and shooting in Ottawa), and the test data contains unseen rumours about natural disasters. The target of the stance, rumour story, is implied by the source tweet that initialises the conversation. Hence we consider the source tweet as the target. Among the four stances, *support* and *deny* classes are the most informative for rumour verification, while the *comment* class is the least useful (Scarton et al. 2020*a*). Therefore, we annotate all the replies in *support*, *deny* and *query* classes in the validation and test sets, with 50 randomly sampled *comments* from each set.

Annotation Process Two annotators manually categorised each reply into either target-independent or -dependent, by answering one question: "*do you think you need the source tweet to infer the stance of this reply?*" Aiming to validate the annotations, annotators were also asked to classify the stance of the tweets. We then compared their assigned class with the gold standard label and, if they differed, we altered their annotation from target-independent to -dependent. Annotators did not have access to the source tweet and the tweets from validation and test sets were shuffled before annotation. The inter-annotator agreement is of 72.5% and Cohen's Kappa is 0.565.

[3]The dataset contains Twitter and Redddit. To alleviate the impact of text length, we focus on the Twitter data only

Result We observe a significant amount of data whose stance can be deduced without knowing the specific rumour story (Table 2.2), especially in the *deny* and *query* classes. More than 50% *denies* are target-independent in the validation and test sets. Target-independent *denies* are tweets that directly cast doubt with negation words (e.g. “Fake news”, "This is false"). The *queries* tend to be target-independent, since most of them are interrogative sentences asking for more evidence. However, the annotators did not identify many of them due to the ambiguity or non-informativeness of the texts (e.g., “blood clot?”, “WHAT?”). Most of the target-independent *supports* are retweets and quote tweets, whose context is self-contained. Tweets in the *comment* class are less relevant to the veracity of the rumour story, however, determining their relevance normally necessitates reasoning with the rumour story itself. We present more examples of target-independent and -dependent tweets in the Appendix 2.6.1.

Dataset	Support	Deny	Query	Comment
Validation	20 (29%)	35 (51%)	42 (40%)	0 (0%)
Test	12 (13%)	66 (72%)	17 (30%)	0 (0%)

Table 2.2: *Number of target-independent tweets in each stance in the validation and test sets (proportion in brackets).*

2.2.2 Model Evaluation

Given a source tweet (s_i), reply tweet to classify (r_i), other replies in the conversation (o_i) and stance label (l_i), we consider two types of supervised models: target-oblivious ($f(r_i) \rightarrow l_i$) and target-aware ($f(s_i, r_i)$ or $f(s_i, r_i, o_i) \rightarrow l_i$) models. We also evaluate a recent large language model (LLM) in zero-shot setting.

Type	Model	Full set	Target-dependent subset			Target-independent subset		
		wF_2	wF_2	$F_2(S)$	$F_2(D)$	$wF2$	$F_2(S)$	$F_2(D)$
Target-oblivious	BERTweet	**0.477**	**0.346**	0.294	0.206	**0.749**	0.615	**0.894**
Target-aware	BERTweet	0.435	0.329	0.313	0.167	0.635	0.464	0.778
	BLCU-NLP	0.371	0.223	0.080	0.217	0.399	0.000	0.737
	BUT-FIT	0.309	0.176	0.020	0.047	0.371	0.102	0.495
	Branch-LSTM	0.150	0.139	0.020	0.048	0.142	0.102	0.056
	Hierarchical-BERT	0.235	0.137	0.065	0.017	0.234	0.017	0.293
LLMs	LLaMA (reply)	0.256	0.227	**0.390**	0.000	0.319	0.417	0.093
	LLaMA (source & reply)	0.419	0.318	0.326	**0.234**	0.685	**0.678**	0.714

Table 2.3: *Model performance over the full test set, target-dependent and -independent direct replies (averaged over experiments.). $F_2(S)$ and $F_2(D)$ denote the F_2 scores over support and deny classes, respectively. Highest performance is in bold, with statistical significance (t test, p value<0.05).*

Experimental Setups

Target-oblivious Models We fine-tune different transformer-based models, whose input is the reply tweet ($f(r_i)$). We present the results using BERTweet (Nguyen et al. 2020)

	Target-dependent subset				Target-independent subset			
	Support	Deny	Query	Comment	Support	Deny	Query	Comment
Mask Source Tweet	40.3	69.9	98.7	85.7	43.0	90.8	98.7	89.0
Shuffle Source Tweet	54.1	82.9	93.6	90.9	57.7	94.9	97.3	83.1

Table 2.4: *The proportion (%) of target-aware BERTweet predictions of direct replies in each class that are not influenced by the masking or shuffling of the source tweets.*

(experiments with BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) achieved similar performance).

Target-aware Models We fine-tune BERTweet, which takes as input both source and reply tweets ($f(s_i, r_i)$). We also evaluate four competitive systems that model the whole conversation thread ($f(s_i, r_i, o_i)$):[4] (1) The winner of the RumourEval 2019 shared task, i.e. *BLCU-NLP* (Yang et al. 2019); (2) *BUT-FIT* (Fajcik et al. 2019), the second place in the 2019 shared task; (3) *Hierarchical-BERT* (Yu et al. 2020), achieving state-of-the-art performance (Hardalov et al. 2022*b*) on the RumourEval 2017 dataset (Derczynski et al. 2017); (4) *Branch-LSTM* (Kochkina et al. 2017), the winner of the RumourEval 2017 shared task and the baseline model for the 2019 task.

LLMs We experiment with the OpenAssistant LLaMA-Based Model (Köpf et al. 2023).[5] We compare the performance between two scenarios: (i) when the source tweet is provided (*LLaMA (source + reply)*) and (ii) when it is not (*LLaMA (reply)*).[6]

Evaluation We adopt the weighted F_2 score proposed by Scarton et al. (2020*a*), which gives higher weights to the *support* and *deny* classes, being more adequate to rumour stance classification.

Results

As shown in Table 2.3, not surprisingly, all the models achieve better results on the target-independent samples, since they normally contain explicit stance-associated words or signals, especially for the *deny* and *query* classes. The target-oblivious model exhibits strong performance over target-independent tweets, indicating that its performance can be attributed to the existence of these samples in the dataset.

We expected that target-aware models, especially the ones that consider the whole conversation information, would perform significantly better than target-oblivious models on the target-dependent tweets for which the context of the source tweet is essential. However, among the target-dependent *supports* and *denies* that the target-oblivious BERTweet couldn't identify, BUT-FIT, Branch-LSTM and Hierarchical-BERT fail to correctly predict any of them as well, casting doubt on the usefulness of these target-aware approaches. BLCU-NLP is the

[4]Performances of these models are lower than the figures reported in their original paper. The reason is that we do not consider the stance of the source tweet towards rumour, mainly belonging to the support class.

[5]https://huggingface.co/OpenAssistant/oasst-sft-6-llama-30b-xor

[6]Due to ethical considerations regarding the exposure of personal data (e.g., to ChatGPT), we opt to use an open-source LLM which was downloaded and hosted on our own server.

only conversation-based system that outperforms the target-oblivious model over the target-dependent *denies*, likely due to their data augmentation for this class. But its performance over the target-dependent *supports* is rather disappointing.

Target-aware BERTweet shows strength on detecting target-dependent *supports*, when compared with its target-oblivious counterpart; however, it falls behind on the *deny* class. The existence of negation words (e.g., "not") in the target-dependent *denies* may contribute to the good generalisation of target-oblivious BERTweet.

LLaMA exhibits competitive results, achieving best performance on the target-dependent samples in the *support* and *deny* classes. However, gaps still exist between the fine-tuned BERTweet models on the full test set. Without the source tweet, the performance drops significantly, except for the target-dependent *supports*.

Target perturbations

Aiming to further investigate the role of the target in target-aware models, we experiment with two perturbations during inference: (1) Masking: the entire source tweet is replaced by a white space; (2) Shuffling: the original source tweet is replaced by a source tweet related to another rumour story so that the reply and "new" source tweets are mismatched. Both approaches should significantly change the model performance over the target-dependent tweets, provided the source tweet is properly reasoned with. We expect the *comment* class to be less impacted because the irrelevance between source and reply tweets should be considered as *comment*. We discuss the results of the target-aware BERTweet, since it is the best performing model in this category (other models showed similar results).

Masking or shuffling the source tweets has minimum impact over the predictions for the *deny*, *query* and *comment* classes (Table 2.4). More than 69% of predictions in each class stay the same, no matter whether the target is essential or not. For the *support* class in which target-aware BERTweet achieves better results over target-dependent samples, 40% to 60% of predictions do not change. The results suggest that target-aware models may be overfitting towards the replies, behaving like a target-oblivious model.

2.3 Ensemble-based Framework

Equipped with the observation of target-independent cases and the lack of reasoning with the target in target-aware models, we propose a simple yet effective ensemble-based framework to leverage the advantage of the target-oblivious model meanwhile improving the performance over the target-dependent samples.

We assume a pre-trained target-oblivious model ($f(r_i;\theta) = p_i$). The aim is to adopt an ensemble with a target-aware model ($f'(s_i, r_i;\theta') = q_i$) where p_i and q_i are posterior probability distribution over the four stance classes for a sample i with a pair of source (s_i) and reply (r_i) tweets. To encourage the target-aware model to learn from target-dependent samples during training, we propose a cross-attention based architecture with a sample re-weight mechanism.

Siamese Network with Cross-attention We utilise a siamese pre-trained transformer-based network (Reimers & Gurevych 2019) to encode the source (s_i) and reply (r_i) tweets.

Then, to explicitly indicate the importance of the tokens in the reply representation (h_{r_i}) with respect to the source representation (h_{s_i}), we calculate the cross-attention (Vaswani et al. 2017) between them, with h_{s_i} as the key and value, and h_{r_i} as the query.

Sample Re-weight We train the model on weighted data, where the weight of instance i is $1 - p_{y_i}$ (p_{y_i} is the posterior probability assigned to the true label y_i) (Clark et al. 2019). The intuition is to encourage the target-aware model to focus on potential target-dependent examples that the target-oblivious model gets wrong.

Implementation Target-oblivious and -aware models are based on BERTweet but our method can be easily generalised to other pre-trained language models. The optimal target-oblivious model is chosen based on the validation set.

2.3.1 Experimental Setup

Datasets We validate our proposed framework on two benchmark datasets: RumourEval 2017 and 2019 datasets.

Comparing Baselines We compare with the Pretext Task-based Hierarchical Contrastive Learning model (*PT-HCL*) (Liang et al. 2022). To the best of our knowledge, PT-HCL is the only study that exploits "target-invariant/-specific features" (Liang et al. 2022) in traditional stance classification. We also present ablations for our proposed method, by removing the sample re-weighting mechanism (*w/o weight*), replacing the cross-attention by self-attention on the concatenation of the source and reply tweet representations (*w/o cross-att*), or both simultaneously (*w/o weight,cross-att*). Performance over RumourEval 2019 dataset is comparable with models in Table 2.5. As for RumourEval 2017, we also compare with its state-of-the-art model (Hierarchical-BERT), target-oblivious and -aware BERTweet and OpenAssistant LLaMa.

2.3.2 Results

As shown in Table 2.5, our proposed approach outperforms PT-HCL on both datasets, also surpassing other models. Removing sample weights or cross-attention would reduce the model performance, indicating their contribution.

We also evaluate our proposed method and its ablations on target-dependent and -independent subsets, as shown in Table 2.6. Comparing with Table 2.3, our model achieves the best results on both target-dependent and -independent examples, confirming that our proposed framework could not only benefit from the target-oblivious model but also enhance the inference between source and reply tweets. Furthermore, ensemble with either cross-attention or sample-weighting based target-aware model could improve the performance on average, if we compare the results of ablations. Sample-weights (w/o cross-att) could guide the target-aware model to focus more on the instances that target-oblivious model struggles with, resulting in more improvement over the target-dependent subsets than the method with only cross-attention (w/o weight). However, the differences are not statistically significant.

Method	2019 dataset	2017 dataset
PT-HCL	0.452	0.431
Hierarchical-BERT	0.235	0.275
LLaMA	0.419	0.314
Target-oblivious BERTweet	0.477	0.425
Target-aware BERTweet	0.435	0.426
Proposed Method	**0.510**	**0.452**
w/o weight	0.458	0.421
w/o cross-att	0.438	0.417
w/o weight,cross-att	0.436	0.419

Table 2.5: *Averaged wF_2 over experiments for two datasets. Highest performance is in bold, with statistical significance between the proposed method (t test, p value <0.05).*

	Target-dependent subset			Target-independent subset		
	wF_2	$F_2(S)$	$F_2(D)$	$wF2$	$F_2(S)$	$F_2(D)$
Proposed Method	**0.396**	**0.399**	**0.211**	**0.802**	**0.732**	**0.901**
w/o weight	0.346	0.326	0.197	0.680	0.532	0.827
w/o cross-att	0.355	0.328	0.210	0.669	0.537	0.802
w/o weight,cross-att	0.314	0.322	0.191	0.627	0.390	0.804

Table 2.6: *$wF2$ scores of our proposed method and ablations over the target-dependent and -independent subsets of RumourEval 2019 test set. Highest performance is in bold, with statistical significance between "w/o weight,cross-att" (t test, p value<0.05).*

2.4 Conclusion

In this paper, we explore the role of the target in rumour stance classification. Our study suggests the strong performance of target-oblivious models could be explained by the existence of target-independent texts in real-world data. We point out the unexpected weakness of the target-aware models and consequently propose a cross-attention based architecture with a sample re-weight mechanism, achieving the best results on two benchmark datasets. We also release our annotations of target-dependent or -independent replies to facilitate future research and model evaluations. Finally, we argue that research in this area (and other textual entailment tasks) should conduct a thorough data analysis in order to fully understand models' performance, going beyond automatic metrics results.

2.5 Acknowledgements

This work is funded by the European Union under action number 2020-EU-IA-0282 and agreement number INEA/CEF/ICT/A2020/2381686 (EDMO Ireland).[7] and by EMIF managed by the Calouste Gulbenkian Foundation[8] under the "Supporting Research into Media, Disinformation and Information Literacy Across Europe" call (ExU – project number: 291191).[9]

[7] `https://edmohub.ie`

[8] The sole responsibility for any content supported by the European Media and Information Fund lies with the author(s) and it may not necessarily reflect the positions of the EMIF and the Fund Partners, the Calouste Gulbenkian Foundation and the European University Institute.

[9] `exuproject.sites.sheffield.ac.uk`

Yue Li is supported by a Sheffield–China Scholarships Council PhD Scholarship.

2.6 Appendix

2.6.1 Target-Independent and Target-Dependent Examples

We present examples of target-independent and -dependent tweets in the RumourEval dataset for different stance classes in Table 2.7.

Stance	Source Tweet	Reply Tweet
Target-dependent replies		
Deny	267 days since Sick Hillary had a press conference.	@USER wha do you mean she had one with Anderson cooper over the telephone
Deny	BREAKING: At least 10 killed in shooting at French satirical newspaper Charlie Hebdo, Paris prosecutor's office says.	@USER 11 killed
Support	Germanwings co-pilot had serious depressive episode: Bild newspaper	@USER The pilot was NOT FIT TO FLY !
Support	Report: Red Cross Was Stealing from Church Doorsteps to Redistribute or Sell Items for Profit?	@USER @USER Stealing is stealing, regardless of how you want to dress it up.
Target-independent replies		
Deny	BREAKING: Illegal Muslim From Iran Arrested For Starting California Wildfire HTTPURL	@USER No source cited in this article, no date... I would not rely on this and neither should you.
Deny	Prince William and Harry donates $ 100 million to Hurricane Harvey Victims – News 360	@USER Fake news!!
Query	Black Lives Matter THUGS Blocking Emergency Crews From Reaching Hurricane Victims via @USER	@USER @USER @USER Where and when ? Other links to ?
Support	Ongoing hostage situation in Sydney café. Major landmarks like the Sydney Opera House evacuated	Special Prayers for tonight "@USER: Ongoing hostage situation in Sydney café."
Support	Mike Pence Disappointed God Has Never Asked Him To Kill One Of Own Children	@USER There's lot of truth in this

Table 2.7: *Examples of target-independent and -dependent tweets*

2.6.2 Pre-processing

User mentions, URLs and emojis are treated in the same way as in the pre-training of BERTweet. Hashtags are removed from the tweets. Most of them are related to the name of the news events rather than the rumour story (the *target*), e.g., #CharlieHebdo. The max sequence length is set to 128.

2.6.3 Training Process

BERTweet We use the `bertweet-base`.[10] During fine-tuning, we employ the transformers library (Wolf et al. 2019) and adopt AdamW (Loshchilov & Hutter 2017). We introduce class weights in the loss function to treat the imbalance data problem. The class weights are computed according to the class distribution of the training data. We use grid search for hyperparameter tuning and the optimal hyperparameters are determined based on the $wF2$ score on the validation set. We search the batch size from [16, 32] and the learning rate from [1e-5,3e-5,5e-5,7e-5,1e-4]. We set the maximum epochs to 50 and use an early stopping

[10] `https://huggingface.co/vinai/bertweet-base`

strategy. The best model checkpoint is selected according to the $wF2$ score on the official validation set. For each model, we repeat the fine-tuning process for five times with different random seeds.

Hierarchical BERT. We adopt the implementation of Yu et al. (2020) for the single task model[11] for rumour stance detection and re-train it with the RumourEval 2019 training set. We use the hyperparameters suggested by the authors. Due to memory limitations, we reduce the batch size from 2 to 1 and tune the learning rate from [1e-5,3e-5,5e-5,7e-5,1e-4]. We repeat the training process for five times with different random seeds.

Branch-LSTM We directly utilise the trained model shared by Kochkina et al. (2017) [12].

Pretext Task-based Hierarchical Contrastive Learning (PT-HCL) We adopt the implementation of Liang et al. (2022) [13]. Following their approaches, we first train an over-fitting target-aware BERTweet-base model, then compare the predictions on the training set before and after removing the *source* tweet. For the reply tweets whose predictions stay the same, we assume they contain "target-invariant features" (Liang et al. 2022). And the rest of tweets include "target-specific features" (Liang et al. 2022). Then we adopt their proposed contrastive loss to learn the correlation and difference between and within "target-invariant and -specific features". Although the authors' implementation does not consider the class imbalance problem, we train a model with class weighted loss function for a fair comparison with our approach (i.e. the value we report in Table 2.5 is with class weighted loss). The $wf2$ score without class weighted loss function is 0.292.

Ensemble-based Approaches We search optimal batch size from [16, 32] and learning rate from [1e-5,3e-5,5e-5,7e-5,1e-4]. Other settings are the same as experiments in fine-tuning the BERTweet.

[11] `https://github.com/jefferyYu/DualHierarchicalTransformer`
[12] `https://github.com/kochkinaelena/RumourEval2019`
[13] `https://github.com/HITSZ-HLT/PT-HCL`

Chapter 3

Publication II: Rumour Stance Classification Adaptation Under Domain Shift: A Case Study of Rumours in Ireland

Overview Publication I examines the generalisability of existing rumour stance classification models. Based on the identified limitation of current systems, we propose a new framework that achieves the best generalisation performance on unseen rumours in RumourEval datasets. However, due to the constraints of available public datasets, current models, including our proposed framework, have not been evaluated under substantial domain shift yet. Consequently, their reliability on real-world application remains in question. To address this gap, the publication presented in this chapter introduces a new rumour stance classification dataset with significant topical, lexical and label shift from current datasets. We then systematically evaluate the model performance drop under this domain shift and propose a new domain adaptation method that does not require human-labelled data from both the source and target domains.

Rumour Stance Classification Adaptation Under Domain Shift: A Case Study of Rumours in Ireland

Yue Li[1], *Cian McGrath*[2], *Kirsty Park*[2], *Eileen Culloty*[2]
Kalina Bontcheva[1], *Carolina Scarton*[1]
[1]School of Computer Science, University of Sheffield, UK
[2]School of Communications, Dublin City University, Ireland

Abstract

Considering a conversation initialised by a rumour-introducing source post on social media, rumour stance classification aims to identify the opinion (e.g. agree or disagree) of its replies towards the source post. Methods for automatic rumour stance classification have yielded high performances; however, none of them have been evaluated on a dataset that is substantially different from the data they were trained and tested on, putting their reliability on real-world applications into question. This paper releases a novel rumour stance classification dataset (ISLES), focusing on online rumours spread in Ireland, that features substantial differences from commonly used datasets over topic, lexicon and stance distribution. We evaluate the top-performing models trained with RumourEval data on ISLES and observe considerable performance degradation when compared to the performance on RumourEval test sets. Further, we seek to adapt a RumourEval-trained model to the domain of rumours spread in Ireland, in the absence of both labelled RumourEval and Irish-domain rumours data. By leveraging synthetically generated Irish-domain rumour stance data with large language models, our proposed self-training framework for domain adaptation successfully boosts the performance of the RumourEval-trained model by nearly 127% on ISLES.

3.1 Introduction

Social media platforms can be used to deliver information more rapidly and freely than traditional media; however, this inevitably leads to a large number of unfiltered, misleading or even malicious posts (Phuvipadawat & Murata 2010, Lazer et al. 2009). Automatic tracking or analysing users' reaction and stance (e.g., whether they agree/disagree) is a useful tool for content verification professionals to identify and verify check-worthy information (Guo et al. 2022, Hardalov et al. 2022*b*, Nakov, Corney, Hasanain, Alam, Elsayed, Barrón-Cedeño, Papotti, Shaar & Da San Martino 2021). Progress has been made in automatic stance classification for rumour (Qazvinian et al. 2011, Zubiaga et al. 2016, Derczynski et al. 2017, Gorrell et al. 2019*a*) or misinformation analysis (Zheng et al. 2022, Hossain et al. 2020, Lavrouk et al. 2024) on social media. In this work, we focus on the rumour stance classification task proposed in the pipeline for rumour analysis on social media (Zubiaga et al. 2018*b*). Given a conversation initialised by a rumour-introducing post (i.e., source post), the aim of this task is to classify the stance of the replies towards the source post into four categories (see examples in Table 3.1):

- *Support*: the author of the reply agrees with the source post;

- *Deny*: the author of the reply disagrees with the source post;
- *Query*: the author of the reply asks for additional evidence about the source post;
- *Comment*: the author of the reply makes their own comment without a clear stance towards the source post.

Source Tweet: Every worker who wants to work in Ireland is welcome as far as I am concerned, but allowing people to travel in a pandemic to pick fruit is not essential, or specialist. It is super exploitation for super profits. No social distancing here. Shame on @USER.
Reply 1: @USER #lies #fakenews nothing worse than a public servant using lies to peddle their agenda while scaring people who are already scared enough. **[Deny]**
Reply 2: @USER So glad you are now in Europe! What a difference you will make! Too bad you still can't be bothered to check facts! **[Deny]**
Reply 3: @USER I don't agree with many things that you say and I would go as far as saying "I don't like you or your policies" but, I 100% agree with your statement. **[Support]**
Reply 4: @USER Surely Keelings needed permission to bring Bulgarians here during pandemic? Who gave them the okay? **[Query]**

Table 3.1: *An example in our novel dataset with the different stances towards a source tweet that initialises a rumour.*

Studies in this area have largely been driven by three annotated corpora: PHEME (Twitter, now X) (Zubiaga et al. 2016), RumourEval 2017 (X) (Derczynski et al. 2017) and RumourEval 2019 (X and Reddit) (Gorrell et al. 2019*a*) datasets. However, the above three datasets have notable overlap. Specifically, the RumourEval 2017 dataset constitutes the English subset of the PHEME dataset. In the RumourEval 2019 dataset, the RumourEval 2017 data serves as the training set, augmented with the Reddit data. The new RumourEval 2019 test set contains conversations about unseen rumours from X and Reddit. Additionally, the above three datasets (considering both training and test splits) share a common imbalanced stance distribution, with the *comment* class substantially outweighing the other three classes. Consequently, there is a risk that the current top-performing models may not perform as well as expected when presented with datasets featuring varying topic domains, particularly those with different stance distributions.

To fill this research gap, we present the novel **I**rish-domain **S**tance c**L**assification t**E**sting **S**et (`ISLES`), covering diverse rumours circulating in Ireland or of interest to Irish citizens: such as Brexit, Irish politics, COVID-19 and Russia-Ukraine war. This novel dataset comprises a total of 1,151 tweets in English, annotated for stance towards the rumourous source tweets, following the same annotation scheme employed in the RumourEval datasets. `ISLES` features topics and rumours significantly different from those in the RumourEval datasets, demonstrating substantial vocabulary differences from both RumourEval 2019 training and test sets, as well as a significantly distinct stance distribution. With ISLES, we can then assess and analyse the generalisability of several top-performing rumour stance classification models under such domain shift, and observe noticeable performance degradation.

Seeking to adapt rumour stance classification models to the domain of rumours spread in Ireland, we innovate by exploring *source-free unsupervised domain adaptation* (Fang et al. 2024) (SFUDA, also known as model adaptation (Li, Jiao, Cao, Wong & Wu 2020)). Unlike unsupervised domain adaptation (UDA), SFUDA restricts access to the source data, aiming to transfer the knowledge from a model pre-trained with the source data (i.e., RumourEval 2019 dataset) to the unlabelled target domain (i.e., Irish-domain rumours). Mainstream UDA methods are not applicable in SFUDA, as they typically work to either align the source and target domain distributions (Li, Chen, Ding, Zhu, Lu & Shen 2020, Long et al. 2015, Cui et al. 2020) or learn domain-invariant representations (Ganin & Lempitsky 2015, Tzeng et al. 2017, Long et al. 2018, Malik et al. 2023, Trung et al. 2021), necessitating the availability of the source data. The study of SFUDA in the field of natural language processing (NLP) is limited, primarily focusing on the medical domain because of privacy regulations. However, in recent years, data accessibility has become increasingly problematic for the computational social science field due to (1) API restrictions imposed by social media platforms like X (e.g., no academic access); (2) reproducibility issues: with the risk of data deletion or removal, most useful posts for training and evaluation may become inaccessible over time, especially if only post IDs are published.

Pseudo labelling is a commonly adopted approach for SFUDA in NLP (Yoon et al. 2021, Kumar et al. 2021, Wang, Wu, Liu & Liu 2021, Kurniawan et al. 2021, Su et al. 2021, 2022), where the target model is initialised by the source model and iteratively updated by the pseudo labels generated for the unlabelled target domain data by the source model. Traditional data augmentation approaches (e.g., lexical substitution) has been utilised on the pseudo-labelled samples (Kumar et al. 2021, Su et al. 2021, 2022), aiming to increase the target data diversity and mitigate the pseudo label noises, but the improvement is relatively marginal (Su et al. 2022). We hypothesise that one of the reasons is that these traditional augmentation methods involve comparatively simple transformations, which may not capture the full range of linguistic variations presented in the target domain or incorporate substantial target domain knowledge. Inspired by the impressive ability of the generative large language models (LLMs) on text generation, data augmentation, knowledge retrieval and contextual understanding (Wei, Bosma, Zhao, Guu, Yu, Lester, Du, Dai & Le 2022, Min et al. 2023, Ouyang et al. 2022, Stacey et al. 2024, Zhang, Tian, Wei, Zeng & Mao 2024, Lan et al. 2024, Zhu et al. 2025), we propose to utilise LLMs to synthetically generate replies with different stances for rumours in the Irish domain. This introduces another novelty over previous work: the labels of our augmented replies are not predicted by the source model which suffers from performance drop in the target domain.

Overall, our contributions in this paper can be summarised as follow:

- To the best of our knowledge, we are the first to establish an **English rumour stance classification test set** (`ISLES`[1]) that **significantly differs from RumourEval datasets** in both vocabulary and stance distribution, making our test set well-suited for evaluation and adaptation under domain shift for this task.
- We assess the **generalisability of top-performing supervised models under domain shift**, also comparing with the zero-shot in-context learning ability of three different LLMs.

[1]Dataset: `https://zenodo.org/records/10830188`

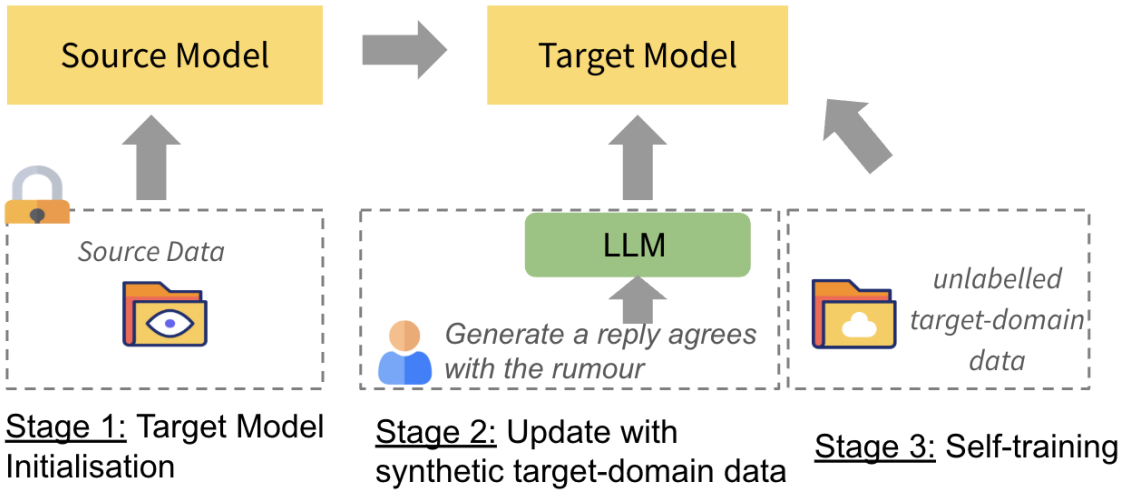


Figure 3.1: *An overview of our proposed framework for source-free unsupervised domain adaptation for rumour stance classification.*

- We explore a **novel source-free unsupervised domain adaptation setting for rumour stance classification** and propose a **new self-training framework integrated with the silver-labelled synthetic data generated by an LLM** (as shown in Figure 3.1). Our approach demonstrates a remarkable improvement in performance on the target Irish domain, with an increase of nearly 35 points in weighted $F2$ score (Scarton et al. 2020*a*) when compared to the model before adaptation.
- We conduct a **comprehensive error analysis** and investigate the potential **impact of the number of unlabelled real-world Irish-domain data and synthetic Irish-domain data** on the model adaptation, providing insights into the data resource allocation and data augmentation strategies.

In the following sections, we first discuss related work in Section 3.2. Section 3.3 and 3.4 introduce our newly established `ISLES` dataset and present benchmarking results on supervised training and in-context learning. Section 3.5 present our proposed method and analysis for source-free unsupervised domain adaptation. Finally, we conclude the paper in Section 3.6.

3.2 Related Work

3.2.1 Rumour Stance Classification

Datasets There are two established English rumour stance classification datasets: RumourEval 2017 (Derczynski et al. 2017) and RumourEval 2019 (Gorrell et al. 2019*a*). RumourEval 2017 dataset (also English PHEME dataset (Zubiaga et al. 2016)) contains a total of 5,568 tweets, covering rumours prompted by five breaking news stories (such as a shooting in Ottawa and a siege in Sydney). These rumours and conversations are collected by monitoring the posts on X during the emergence of the breaking news and then manually identified by journalists. It also contains four known rumours such as "Putin was missing". The X

conversations are then sampled according to the number of retweets of the source tweet. RumourEval 2019 dataset extends the RumourEval 2017 dataset by introducing a new X test set (1,066 tweets) about rumours related to natural disasters and a set of Reddit data (1,940 posts). As for tweets, the rumours were retrieved from fact-checking websites (Snopes.com and Politifact.com), and the conversations were collected by querying the Twitter API. The Reddit posts were manually selected from debunking forums. A comprehensive comparison between the RumourEval 2019 Twitter dataset and `ISLES` can be found in section 3.3.4. It is worth mentioning that three non-English rumour stance classification datasets have also been constructed for Russian (Lozhnikov et al. 2018), Danish (Lillie et al. 2019*a*) and German (Zubiaga et al. 2016).

Another three closely related stance classification X datasets are CovidLies (Hossain et al. 2020), STANCEOSAURUS (Zheng et al. 2022) and STANCEOSAURUS 2.0 (Lavrouk et al. 2024), where tweets are annotated for the stance towards a given claim, rather than a source tweet which implies a rumour story. They also adopt different annotation schemes, with three-way or five-way stance classes.

Another line of research associated with rumour stance classification is framing stance classification as fact-checking, such as the FEVER-like datasets (Thorne et al. 2018, Aly et al. 2021, Wadden et al. 2020, Saakyan et al. 2021). The datasets typically consist of a pair of artificial, or real-world claim and evidence. The stance of the evidence towards the claim could be seen as the veracity of the claim. Although formed as stance classification, the nature of these stance classification tasks is also connected with natural language inference (Hardalov et al. 2022*b*).

Automatic Rumour Stance Classification Earlier work on rumour stance classification mainly focuses on feature engineering (Aker et al. 2017, Ghanem et al. 2019, García Lozano et al. 2017, Bahuleyan & Vechtomova 2017), modelling the X's conversation structure (e.g., linear sequences (Kochkina et al. 2017, Li, Sujana & Kao 2020, Yang et al. 2019, Fajcik et al. 2019), tree (Kumar & Carley 2019) and graph (Li, Sujana & Kao 2020, Wei et al. 2019)), and jointly learning with rumour veracity (Kochkina et al. 2018, Kumar & Carley 2019, Wei et al. 2019, Li, Sujana & Kao 2020). Another challenge in this task is imbalanced stance distribution. The two most informative stance classes (support and deny classes) are the minority classes in the RumourEval datasets, while the least useful comment stance is the majority class. So efforts have also been made to investigate a proper imbalanced data treatment (Li & Scarton 2020) and evaluation metrics (Scarton et al. 2020*a*) for this task. Overall, the methods focus on supervised approaches, progressing from feature-based models (e.g., Aker et al. 2017), neural approaches (e.g., Kochkina et al. 2017, Kumar & Carley 2019), to fine-tuning pre-trained language models (PLMs) (e.g., Yang et al. 2019, Li, Sujana & Kao 2020). The approaches are all trained and evaluated on RumourEval datasets, whose generalisability to a dataset with significantly different topics, lexicons and stance distribution is under-explored.

Recently, LLMs are increasingly used for classification tasks via zero-shot in-context learning, where models are prompted to select an option from a pre-defined set of classification labels Wang et al. (2022), Antypas et al. (2023*a*), Mu et al. (2024). Social media datasets have been benchmarked on a wide range of LLMs (Antypas et al. 2023*b*), but rumour stance classification seems to be lagging behind in this direction probably due to the relatively small scale

of datasets and research community. LLMs have also been applied to assist the supervised training of the PLMs for stance classification, such as chain-of-thought embedding generation (Gatto et al. 2023), contextual knowledge extraction (Zhang, Tian, Wei, Zeng & Mao 2024, Zhang, Li, Zhang & Xu 2024, Zhu et al. 2025) and data augmentation (Wagner et al. 2024). However, the approaches focus on generic stance classification, whose task formulation and evaluation are distinct from rumour stance classification (Hardalov et al. 2022*b*, Scarton et al. 2020*a*, Li & Scarton 2024).

Model Generalisability Evaluation and Model Adaptation As far as we know, there are no previous studies on generalisability evaluation for rumour stance classification models. Zheng et al. (2022) assess the performance drop when the model is trained with the STANCEOSAURUS data and evaluated on the RumourEval 2019 dataset (and vice versa). However, since the stance classes of the two datasets can not be directly matched, the *query* and *comment* classes in the RumourEval dataset are merged.

Supervised domain adaptation method (i.e., EasyAdapt (Daumé III 2007, Bai et al. 2021)) is adopted to improve the performance over both STANCEOSAURUS and RumourEval 2019 datasets, where both labelled source and target data are utilised. Unsupervised adaptation methods have also been explored, aiming to transfer knowledge from multiple stance classification datasets to out-of-domain unseen stance classification datasets with potential different stance definitions (Hardalov et al. 2021, Schiller et al. 2021, Li et al. 2021, Arakelyan et al. 2023). However, a key difference between their studies and ours is that we do not assume the access to the source data (i.e. the multiple stance classification datasets in their work), making their approaches inapplicable in our adaptation scenario.

3.2.2 Source-Free Unsupervised Domain Adaptation

The main challenge of SFUDA is to learn the target domain representation without ground truth annotations, only relying on a source-data trained classifier as a proxy for the knowledge of the source domain. A growing number of approaches have been proposed to address this challenge in computer vision, including re-framing into UDA by generating images (Kurmi et al. 2021, Yang et al. 2022, Hong et al. 2022, Tian, Ma, Zhang, Peng & Xue 2021, Li, Jiao, Cao, Wong & Wu 2020) or distributions (Tian, Zhang, Li & Xu 2021, Ding et al. 2022) for the source domain, self-training (i.e., pseudo labelling) with label refinement (Liang et al. 2020, Kim et al. 2021, Chen et al. 2022), jointly training with a self-supervised auxiliary task (e.g., contrastive learning) (Chen et al. 2022, Sun et al. 2020) and domain alignment via statistics (Klingner et al. 2022, Liu et al. 2021).

The few studies in NLP mainly focus on tasks in clinical domain (i.e., negation detection and time expression recognition driven by the SemEval 2021 shared Task 10 (Laparra et al. 2021)), as well as aspect-based sentiment analysis (Zhao et al. 2024). The most commonly used approach is pseudo labelling, where a source model is iteratively updated by pseudo labels it generated for the unlabelled target data (Yoon et al. 2021, Kumar et al. 2021, Wang, Wu, Liu & Liu 2021, Kurniawan et al. 2021, Su et al. 2021). These kind of models primarily concentrate on strategies for the selection or refinement of the high-quality pseudo labels based on the prediction probability (Su et al. 2021, Kurniawan et al. 2021), prediction entropy (Yoon et al. 2021, Kumar et al. 2021, Wang, Wu, Liu & Liu 2021), or clustering (Zhao et al. 2024). Active learning tends to show more strength than pseudo labelling in SFUDA,

where a source model is utilised to select a small subset of examples in the target domain to be manually annotated (Su et al. 2022). The objective is to choose examples that are either the most informative or the most uncertain to the source model, aiming to enhance the model's performance over the target domain with minimal labelling effort. However, the effectiveness of active learning could be hindered in cases where the annotation schema is complex or the classification task is inherently subjective (Su et al. 2021).

Lexical-substitution-based data augmentation (Miao et al. 2020) for the target data has also been considered to combine with pseudo labelling and actively learning, aiming to increase target data diversity or alleviate pseudo labels' noise (Su et al. 2022, Kumar et al. 2021, Su et al. 2021). They follow the same workflow, where data augmentation is applied on the selected target data with high-quality pseudo labels (for pseudo labelling) or with human annotations (for actively learning). However, in terms of pseudo labelling, the benefit of such framework may be limited due to: (1) limited target data diversity, where the traditional data augmentation methods (such as lexical substitution (Miao et al. 2020, Arefyev et al. 2020, Zhou et al. 2019), back-translation (Sennrich et al. 2016, Wang, Yin, Lin & Xiong 2021), or noise injection (Wei & Zou 2019)) involve relatively simple transformations, which may not capture substantial linguistic variations present in the target domain or may introduce considerable amount of domain knowledge; (2) the pseudo labels are generated by the model trained with the source data, assuming the source and target data share certain levels of similarity that can be captured and transferred by the model. However, when there is significant discrepancy between the source and target domains, the model's generalisability may not be sufficient for generating useful pseudo labels.

3.3 ISLES: Irish-domain Stance cLassification tEsting Set

In this section, we present ISLES comprising of X's conversations initialised by rumours that are of interest to Irish citizens. We explore rumours spread in Ireland since this is an unexplored domain, presenting significantly different rumours from those spread in other English-speaking countries (e.g., UK and US). This way, we could guarantee significant data diversity, while still using the same language (experiments with model adaptation for languages other than English is left for future work). We design the data sampling process to ensure topic diversity and low vocabulary overlapping with the RumourEval 2019 dataset. The reply posts are manually annotated for stance (i.e. *support*, *deny*, *query*, *comment*) towards the rumour-introducing source tweet, following the annotation scheme of the RumourEval datasets (Derczynski et al. 2017, Gorrell et al. 2019*a*). At the end of this section, we also present a systematic comparison to the RumourEval 2019 dataset.

3.3.1 Data Collection

We collect fact-checked rumours from *The Journal FactCheck*[2] (a verified Irish signatory of the *International Fact-Checking Network*[3]). We retrieve all rumours, together with any embedded tweets from the published fact-checking articles on their website from February 2016 to July 2022. Following the definition proposed by Jiang, Song, Scarton, Singh, Aker & Bontcheva

[2] https://www.thejournal.ie/factcheck/news/

[3] https://ifcncodeofprinciples.poynter.org/signatories

(2023), we then manually identify and select the tweets that deliver the same rumour fact-checked in the articles. We filter out two types of rumourous tweets: (1) cases where rumours are only spread by images or videos without any informative textual descriptions; (2) the number of replies is less than ten. This results in 14 distinct rumours and 16 corresponding rumour-introducing source tweets. The rumours cover topics specifically of interest to Irish citizens (e.g., Irish politics and Brexit) as well as more general topics (e.g., COVID-19 and Russia-Ukraine war). Next, we collect the whole conversation thread initialised by the 16 source tweets through the X API,[4] yielding a total of 11,048 reply tweets. We list the rumours, their corresponding veracity and the number of replies in Appendix 3.7.1.

3.3.2 Data Filtering and Sampling

Data Filtering We filter out two types of replies to include more informative textual tweets in our dataset: (1) replies that only contain URLs, images, videos, emojis, user mentions or any combination of these; (2) since responses in long conversation become less informative and/or gradually shift towards topics less relevant to the rumour (Kochkina & Liakata 2020), we only annotate direct replies and those indirect replies that respond to the direct replies. This leaves us a total of 6,017 replies.

Data Sampling The aim is for `ISLES` to have a comparable size to that of RumourEval 2019 test set yet with substantial divergence in linguistic characteristics, in order to facilitate the evaluation of models in a domain shift setting. Therefore, we employ vocabulary overlap as a measure of similarity (Gururangan et al. 2020) and sample more tweets from the conversations that are less similar to those in RumourEval 2019.

Specifically, we categorise the rumours in `ISLES` into five groups: Brexit, Irish politics, Russia-Ukraine war, COVID-19 and Others. As data pre-processing, we perform lemmatisation and remove the stop words, punctuation, user mentions, URLs, emojis and hashtags.[5] We then calculate the vocabulary overlap between the RumourEval 2019 dataset and each group. The groups are then sorted in ascending order of similarity: Brexit, Irish politics, COVID-19, Others and Russia-Ukraine war. We proportionately sample a total of 1,151 tweets[6] from each group based on the normalised ratio of vocabulary overlap in the five groups (figures can be found in Appendix 3.7.1). Within each group, we sample tweets equally from each conversation while preserving the distribution of direct and indirect replies. In the cases when the conversation contains fewer replies than the allocated figure, the remaining numbers are distributed to other conversations in the same group.

3.3.3 Stance Annotation

We follow the same annotation scheme as the RumourEval datasets proposed by Zubiaga et al. (2015). We first split the nested conversations into 2-tuples and 3-tuples for direct and indirect replies, respectively. The 3-tuples consists of the reply tweet to be annotated, the rumourous source tweet it indirectly replies to, and the parent tweet it directly responds to. Similarly,

[4] `https://developer.x.com/en/docs/x-api`

[5] We use the English stop words list, WordNetLemmatizer and TweetTokenizer implemented on NLTK: `https://www.nltk.org/`

[6] We start from 1,000 tweets and increase the number of tweets during annotation.

the 2-tuples contain the source tweet and the direct reply. User mentions are anonymised by replacing them with a generic identifier (@USER), ensuring the privacy and confidentiality of individual users. During annotation, annotators are asked to read the 2- or 3-tuples of tweets and determine the stance of the reply tweet towards the source tweet. For indirect replies, the stance towards their parent tweets are also annotated. The annotation is carried out on GATE Teamware (Wilby et al. 2023)[7], a collaborative web-based annotation tool. We present the annotation guidelines and the interface of the annotation tool in Appendix 3.7.2.

Following prior work (Mu et al. 2023, Lillie et al. 2019*b*), each tweet is annotated by at least two annotators. We resolve the annotation conflict by involving an extra annotator. If all three annotators disagree, a fourth annotator is added. We remove tweets for which all four annotators could not reach an agreement. Thus, the maximum number of annotators for each tweet is four. During annotation, each annotator is also given an additional 30 tweets that have been annotated by three Irish researchers in our team. We use the agreement level between each annotator and our researchers as a measure for quality control. The data points annotated by low-quality annotators are then re-annotated by a different annotator. The overall inter-annotator agreement is 51.18%. In comparison, the inter-annotator agreement of RumourEval 2017 and RumourEval 2019 datasets (annotated via crowd-sourcing) are 62.2% and 76.2%, respectively. However, it's worth noting that the RumourEval datasets are significantly imbalanced and dominated by the comment class (see Table 3.2), which we hypothesise to lead to higher agreement among annotators.

The annotators are research students majoring in politics, journalism or NLP in Ireland and the United Kingdom and they worked voluntarily. Before annotation, they completed an online tutorial which introduced our annotation guidelines and annotation tool, together with discussions about edges cases (e.g., instances or scenarios that are at the boundaries or extremes of what the guideline covers). The students were given Amazon gift vouchers as a gratification for their voluntary work (approximately 0.3 pounds per tweet).

3.3.4 Comparison with RumourEval Dataset

Basic Statistics We compare dataset size, class distribution, text length and the number of unique words between `ISLES` and the RumourEval 2019 dataset. We transform user mentions, URLs, and hashtags into generic tokens (@USER, URL, #HASHTAG) in each tweet. As for the calculation of the unique words, we follow the same pre-processing as the one for calculating vocabulary overlap.

The figures are shown in Table 3.2. `ISLES` is more lexically diverse than RumourEval 2019 test set, containing more unique words, despite having less rumours. The median text length of `ISLES` is longer than that of RumourEval data, although approximately 23% posts in RumourEval 2019 are collected from Reddit which contains long sequences. The main reason is that their tweets were collected when the Twitter API had the limitation of 280 characters in the text field. Another clear difference is stance distribution. The *comment* class dominates RumourEval 2019 dataset, with *support* and *deny* as the minority classes. However, the *deny* and *comment* are the two majority classes in `ISLES`. We hypothesise that the change of stance distribution could be explained by the nature of the rumours included in the datasets. For instance, the RumourEval 2019 dataset mainly consists of rumours which emerged during

[7] `https://annotate.gate.ac.uk/`

breaking news (e.g., shooting and plane crash) or natural disasters; whilst rumours in ISLES address long-term events (e.g., Brexit and Russian-Ukraine war). Another potential reason is that ISLES is dominated by source tweets spreading false rumours, attracting more responses that deny given source tweets. For example, nearly 10% replies *deny* the false rumours in the RumourEval training set, while only 6% posts disagree with the true rumours.

		RumourEval 2019		ISLES
		Train	**Test**	**Test**
Stance Distribution	Support	925 (18%)	157 (9%)	261 (23%)
	Deny	378 (7%)	101 (6%)	410 (35%)
	Query	395 (8%)	93 (5%)	80 (7%)
	Comment	3,519 (67%)	1476 (81%)	400 (35%)
Text Length	Min	1	1	37
	Median	17	14	33
	Max	760	1,324	67
# Source Posts		327	81	16
# Reply Posts		4,890	1,746	1,151
# of Unique Words		8,287	2,824	3,931

Table 3.2: *Statistics of RumourEval 2019 datasets and ISLES.*

Dataset Distance Following prior work (Kochkina et al. 2023, Peinelt et al. 2019), we use Jaccard Index and DICE coefficient to measure the distance between RumourEval 2019 dataset and ISLES. Jaccard Index and DICE coefficient (DICE) are defined as follows:

$$\text{Jaccard Index} = \frac{V_a \cap V_b}{V_a \cup V_b} \tag{3.1}$$

$$\text{DICE coefficient} = \frac{2 \times V_a \cap V_b}{|V_a| + |V_b|} \tag{3.2}$$

where V_a and V_b denote the sets of unique words in dataset a and b respectively, and $|V|$ represent the size of the set V. The two metrics, reflecting vocabulary difference, equal to one when the datasets are identical, and zero for datasets with no vocabulary overlap. Results in Table 3.3 show that ISLES contains a considerable number of words that are not covered in the RumourEval 2019 training and test sets. Moreover, the RumourEval 2019 test set and ISLES also exhibit a high difference, further validating our attempt to establish a new test set in a different domain.

3.4 Evaluating Rumour Stance Classification Models

3.4.1 Models

Rumour stance classification models evaluated under domain shift are presented here. The supervised models are selected among the top-performing rumour stance classification models

Dataset Pairs	Jaccard Index	DICE Coefficient
RumourEval 2019 train & ISLES	0.216	0.355
RumourEval 2019 test & ISLES	0.195	0.326

Table 3.3: *The Jaccard Index and DICE coefficient scores between RumourEval 2019 dataset and ISLES.*

with publicly available code that enables reproducibility. We train the models using the official RumourEval 2019 training data, and evaluate their performance over both RumourEval 2019 and ISLES. We compare different types of models in terms of whether they make use of (1) task-specific features; (2) representation of the conversation structure; (3) language models pre-trained with social media data; (4) ensemble techniques. We summarise the properties of the supervised models in Table 3.4. We also evaluate the zero-shot in-context learning ability on three LLMs with different architectures.

Model Name	Feature engineering	Post Representation	Conversation Representation	In-domain pre-training	Ensemble
Ensemble-RUS	Yes	GloVe	Individual posts	No	Yes
branch-LSTM	Yes	word2vec	Linear sequences	No	No
Hierarchical BERT	No	BERT-base	Linear sequences	No	No
Ensemble-Target	No	BERTweet-base	Individual posts	Yes	Yes
RoBERTa (FT & PT)	No	RoBERTa-large	Individual posts	No	No
BERTweet (FT & PT)	No	BERTweet-large	Individual posts	Yes	No

Table 3.4: *Summary of supervised model properties.*

Ensemble-RUS (Aker et al. 2017, Li & Scarton 2020) is an ensemble of logistic regression, random forest (Breiman 2001) and multi-layer perceptron with randomly under-sampling the majority class (i.e., comment class) during training. It adopts task-specific features including semantic (e.g., similarity between the reply tweet and the averaged embedding of a list of stance-related verbs) and tweets' meta data (e.g., count of tweet favourite). The source and reply tweets are represented as the concatenation of their averaged GloVe word embeddings (Pennington et al. 2014).

branch-LSTM Kochkina et al. (2017) splits the tree-structured conversations into linear branches and use Long Short-Term Memory (LSTM) to process them. Each post is represented as word2vec embeddings (Mikolov et al. 2013) plus semantic features.

Hierarchical BERT (Yu et al. 2020) represents the conversation thread as the concatenation of the posts in chronological order. Each post is first encoded by the pre-trained BERT for local representation, then passed through an additional BERT layer to obtain a global representation for the thread.

Ensemble-Target (Li & Scarton 2024) is an ensemble of target-oblivious and -aware models based on small-sized BERTweet. The target-oblivious model takes as input only the reply posts, assuming that the stance of such posts can be inferred without the source tweet. The target-aware model encode the source and reply tweets with a siamese BERTweet network, and then use a cross-attention layer to capture the reasoning between them. The target-aware model is trained with sample weighted data – where the weight of each instance is the posterior probability of the target-oblivious model assigned to the true label.

RoBERTa and BERTweet (Liu et al. 2019, Nguyen et al. 2020) We experiment with the large version of pre-trained RoBERTa and BERTweet models. The BERTweet model is a RoBERTa model pre-trained with Twitter/X data. We consider two methods to adapt the two pre-trained language models (PLMs) on our target task: fine-tuning (FT) and prompt-tuning (PT). Fine-tuning updates all of the parameters of the PLMs, while prompt-tuning prepends trainable continuous prompts to the model input for every layer of the PLM and only optimises the parameters of the prompts during training (Liu, Ji, Fu, Tam, Du, Yang & Tang 2022).

LLMs We examine the zero-shot in-context learning ability of the Llama (Touvron et al. 2023), Mistral (Jiang, Sablayrolles, Mensch, Bamford, Chaplot, Casas, Bressand, Lengyel, Lample, Saulnier et al. 2023) and DeepSeek (Guo et al. 2025) model families. Specifically, we experiment with instruction-tuned Llama-2 with 7-billion (Llama-2 (7b))[8] and 13-billion (Llama-2 (13b))[9] model parameters, Llama-3 with 8 billion parameters (Llama-3 (8b))[10], Mistral with 7-billion parameters (Mistral (7b))[11] and Mixtral with 47-billion parameters (Mixtral (47b))[12], and DeepSeek with 7-billion parameters[13]. Mixtral (47b) is a sparse mixture of eight expert Mistral 7-billion models.

3.4.2 Experimental Setups

For supervised models, we follow the configuration proposed in previous work or keep the original model parameters if there is no further explanation (see Appendix 3.7.3). Models are trained with RumourEval 2019 training set. As for in-context learning of LLMs, we test different prompt templates (see Appendix 3.7.3) and choose the optimal one based on its performance on the validation set. All models are evaluated on RumourEval 2019 and `ISLES`.

[8]https://huggingface.co/meta-Llama/Llama-2-7b-chat
[9]https://huggingface.co/meta-Llama/Llama-2-13b-chat
[10]https://huggingface.co/meta-Llama/Meta-Llama-3-8B-Instruct
[11]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1
[12]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
[13]https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat

	RumourEval 2019 test set			ISLES		
Model Name	$mF1$	GMR	$wF2$	$mF1$	GMR	$wF2$
Supervised Models						
Ensemble-RUS	0.4285	0.5038	0.3677	0.3164	0.1164	0.2171
branch-LSTM	0.4928	0.2774	0.2867	0.2937	0.0000	0.1339
Hierarchical BERT	0.5029	0.4016	0.3545	0.2523	0.1483	0.1297
Ensemble-Target	0.4761	0.6199	0.4833	**0.4473**	**0.4212**	**0.3907**
RoBERTa (FT)	0.5185	0.6124	0.4898	0.4182	0.3282	0.3387
RoBERTa (PT)	0.5269	**0.6270**	0.5028	0.3989	0.2986	0.3039
BERTweet (FT)	**0.5455**	0.6260	**0.5122**	0.4025	0.2904	0.3104
BERTweet (PT)	0.5031	0.6230	0.4930	0.3627	0.2924	0.2807
Zero-shot In-Context Learning						
Llama-2 (7b)	0.2585	0.4125	0.3532	0.2787	0.3065	0.3199
Llama-2 (13b)	0.3004	0.4846	0.3669	0.3642	0.3624	0.4312
Llama-3 (8b)	0.2429	0.3779	0.3050	0.3934	0.3375	**0.4777**
Mistral (7b)	0.2506	0.3716	0.2611	0.3341	0.3336	0.3388
Mixtral (47b)	0.4238	0.4266	0.3662	0.4246	0.4101	0.3601
DeepSeek (7b)	0.4011	0.3848	0.3235	0.3315	0.2853	0.2425

Table 3.5: *Performance of the rumour stance classification models evaluated on RumourEval 2019 test set and ISLES. The best performance is in bold and the lowest performance is underlined.*

We adopt three evaluation metrics: weighted $F2$ ($wF2$) (Scarton et al. 2020*a*),[14] geometric mean of recall (GMR) and macro-$F1$ ($mF1$) (official metric for the RumourEval 2019 shared task). While macro-$F1$ is widely used in NLP evaluation, $wF2$ is a more robust metric for this task, as it gives more weight on the more informative *support* and *deny* classes. GMR is another useful metric for evaluation on imbalanced dataset, as it penalises a model that achieves low performance on a given class.

3.4.3 Results

Supervised Models The results are presented in Table 3.5. As for the supervised models that are trained with the RumourEval 2019 training data, the model performances substantially degrade when evaluated on the out-of-domain ISLES. Also, the ranking of the model performance evaluated on RumourEval 2019 test set does not align with the ranking of these models when evaluated on ISLES. For example, the fine-tuned BERTweet (BERTweet(FT)) achieves the best performance on RumourEval 2019 test set considering $wF2$, however, its performance is eight points lower than Ensemble-Target, the best-performing supervised model on ISLES. In contrast, the Ensemble-Target model only ranks 5th among the eight supervised models tested on RumourEval 2019 test set, in terms of $wF2$. Therefore, developing

[14] We use the same weights as Scarton et al. (2020*a*): *deny* = *support* = 0.40, *query* = 0.15 and *comment* = 0.05.

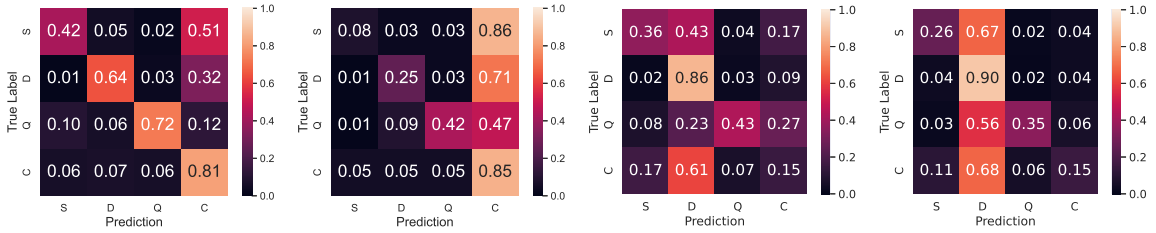
Model Name	Recall				Precision			
	S	D	Q	C	S	D	Q	C
Ensemble-RUS	-96.6	-40.3	-10.7	+18.0	-5.1	+479.4	+11.9	-57.5
branch-LSTM	-100.0	+2412.6	-100.0	-99.4	-100.0	+7.2	-100.0	-74.1
Hierarchical BERT	-89.9	-76.2	-3.2	-3.2	-32.2	-56.0	-23.6	-48.6
Ensemble-Target	-66.4	-26.0	-16.1	+4.2	+83.1	+169.7	+24.7	-53.8
RoBERTa (FT)	-82.5	-31.3	-16.1	-0.1	+16.0	+158.7	+24.1	-37.6
RoBERTa (PT)	-86.6	-49.0	-17.7	+4.8	+36.1	+145.6	+21.4	-36.9
BERTweet (FT)	-83.3	-39.8	-22.5	+3.0	+31.5	+113.3	+5.3	-41.9
BERTweet (PT)	-77.5	-55.5	-35.3	+5.0	+64.8	+144.2	+5.5	-42.4

Table 3.6: *The relative performance change (%) over each stance for the supervised models. Abbreviations: "S", "D", "Q", "C" represents support, deny, query and comment class respectively.*

static models and evaluating them considering only a single test set may be not enough to ensure the creation of generalisable approaches to rumour stance classification for real-world applications, or to avoid over-estimating the model performances.

Furthermore, we observe that approaches that model the conversation structure suffer from poor generalisability to direct and first-level indirect replies in `ISLES`, comparing with the systems that only take the pair of source and reply posts into account. For instance, branch-LSTM could not correctly identify any *supports* in `ISLES`, resulting in a GMR score of zero. Pre-trained contextual embeddings (BERTweet and RoBERTa) show better performance on both test sets than pre-trained word embeddings with task-specific features (Ensemble-RUS and branch-LSTM). Ensemble-Target exhibits the lowest performance drop from RumourEval 2019 to `ISLES`, suggesting good transferability of the learnt rumour-invariant representations. Prompt-tuned PLMs show comparable or even better performance on RumourEval 2019 test set than their fine-tuned versions, however, they also present a relatively higher performance drop than their fine-tuned counterparts when evaluated on `ISLES`.

Finally, we analyse the performance change from RumourEval 2019 test set to `ISLES` for each stance class. We calculate the recall and precision scores for each stance and report the relative differences (i.e., $\frac{(\text{Metric}_{\text{ISLES}} - \text{Metric}_{\text{RumourEval}})}{\text{Metric}_{\text{RumourEval}}}$) for each model in Table 3.6. When evaluated on `ISLES`, all of the models show a decrease in recall of the *support*, *deny*, and *query* classes, especially the *support* class. The branch-LSTM model is an exception, as its predictions become heavily skewed towards the *deny* class on `ISLES`, leading to an increase in recall over the *deny* class. Another interesting observation pertains to the potential impact of imbalanced data treatment during model training. As for the six models that utilise either a class-weighted loss function (RoBERTa (FT & PT), BERTweet (FT & PT) and Ensemble-Target) or random under-sampling (Ensemble-RUS), there is an improvement in the precision scores for the *support*, *deny* and *query* classes (the minority classes in RumouEval 2019 training set) when they are evaluated on `ISLES`. One of the reasons can be that the imbalanced data treatment leads to a model that over-predicts the minority classes in the RumourEval 2019 dataset (i.e., low precision), which may conversely benefit their identification of these classes on `ISLES`.



(a) **(b)** **(c)** **(d)**

Figure 3.2: *Confusion matrices of BERTweet (FT) and Llama-3 (8b) (zero-shot in-context learning) for RumourEval 2019 test set and* `ISLES`*. (a) BERTweet (FT) evaluated on RumourEval 2019 test set; (b) BERTweet (FT) evaluated on* `ISLES`*; (c) Llama-3 (8b) evaluated on RumourEval 2019 test set; (4) Llama-3 (8b) evaluated on* `ISLES`*.*

Zero-shot In-context Learning As shown in Table 3.5 for the RumourEval 2019 test set, when comparing the performance of LLMs in a zero-shot setting to that of small/medium-sized PLMs (fine-tuned/prompt-tuned on RumourEval 2019 training data), LLMs fall behind significantly. However, when evaluated on `ISLES`, the LLMs could achieve comparable or even better results, especially Llama-3 (8b), which has the highest $wF2$ score on `ISLES` among all models. These result suggests the usefulness of the LLMs for rumour stance classification when labelled in-domain data is not available for training.

In contrast to supervised models that tend to incorrectly predict *supports* or *denies* as *comments* (despite imbalance data treatment), LLMs (Llama-2 (7b, 13b), Llama-3 (8b) and Mistral (7b)) often misclassify *comments* as *support*, *deny* or *query* classes. We present confusion matrices for BERTweet (FT) and Llama-3 (8b) in Figure 3.2 as examples. This observation suggests that distinguishing the *comment* class poses a greater challenge for the LLMs, while appropriately integrating with LLMs may present an opportunity to boost the accuracy of the *support* and *deny* classes for the supervised models. Nevertheless, the performances of Mixtral (47b) and DeepSeek (7b) more closely resembles the supervised models, with a clear over-prediction towards the *comment* class.

3.5 Source-Free Unsupervised Domain Adaptation for Rumour Stance Classification

In this section, we explore the SFUDA setting for rumour stance classification, aiming to adapt the model trained on the RumourEval 2019 dataset (i.e., source domain) to the domain of `ISLES` (i.e., target domain). We assume the access to a source model pre-trained with RumourEval 2019 training data and a set of unlabelled rumourous source tweets spread in Ireland (together with their replies). Following Laparra et al. (2021), we adopt a general architecture for the source model by fine-tuning a PLM. However, our self-training framework can be easily generalised to other model architectures.

3.5.1 Methodology

Synthetic Data Generation

Data Generation For each source tweet in the unlabelled set of rumours spread in Ireland, we prompt a generative LLM to synthetically generate N different reply tweets for each stance. The stance defined in each prompt is then used as the silver label of the generated replies. We experiment with Llama-2 (13b), since its $wF2$ score over the RumourEval 2019 validation set in zero-shot in-context learning is the highest, compared to other LLMs. Under the consideration of data efficiency, we opt to generate ten replies per stance per source tweet. The influence of the number of generations on model adaptation is discussed in section 3.5.3. We design the prompt template with the source tweets in the RumourEval 2019 validation set, and use the following input prompt for each stance as:

- **Support:** `Given a source tweet, generate {N} different replies that agree with the source tweet. Avoid using emojis and explicitly using the word agree in the reply. The source tweet is: {source tweet}`
- **Deny:** `Given a source tweet, generate {N} different replies that disagree with the source tweet. Avoid using emojis and explicitly using the word ''disagree'' in the reply. The source tweet is: {source tweet}`
- **Query:** `Given a source tweet, generate {N} different replies that ask for more evidence regarding the source tweet. Avoid emojis. The source tweet is: {source tweet}`
- **Comment:** `Given a source tweet, generate {N} different replies without a clear stance towards the source tweet. Avoid emojis. The source tweet is: {source tweet}`

Particularly, we design the above prompts to avoid introducing learning shortcuts for certain stance classes. Since we use the words "agree" and "disagree" to instruct Llama-2, we observe that the model tends to generate " I completely agree!/Absolutely agree!/I strongly disagree with the idea/I completely disagree with" at the beginning of the *support* and *deny* replies. Thus, we discourage the model to utilise these two words in the replies. We find that it can potentially encourage the model to support or deny the source tweets with lexicons and knowledge in the target domain, which could further benefit the adaptation. Llama-2 tends to generate significantly more emojis in the replies for the *comment* class, which may be explained by the fact that it is relatively more difficult for the model to generate replies without a clear stance. Therefore, we require the model to avoid emojis in the generations. Table 3.7 presents examples of generated replies in different stances for a source tweet.

Data Evaluation Ideally, for each source tweet, the generated replies should be diverse and express the expected stance. For this purpose, we adopt two metrics to evaluate the quality of the generations: selfBLEU (Shu et al. 2019) and the agreement between the silver labels and human judgements. selfBLEU measures the similarity of the N different replies LLM generated for each source tweet in each stance. We report the averaged selfBLEU score

Stance (silver label)	Generated Reply
Support	The potential economic fallout of a no-deal Brexit on Ireland is alarming. The country's heavy reliance on UK trade makes it vulnerable to severe disruption.
	The EU's intransigence on Brexit has put Ireland in a precarious position. A no-deal Brexit would be a disaster for the country's economy.
Deny	Unemployment in German car making cities is already a concern, but it's not solely due to Brexit. The industry has been facing challenges from changing consumer preferences and technological advancements.
	The idea that the EU is "unelected" is a misleading characterization. The EU is made up of democratically elected representatives from its member states, and its institutions are accountable to the people.
Query	Can you provide any specific examples or data to support the claim that a no-deal Brexit would cause 'mayhem' in Northern France? I'd like to see some evidence before accepting this assertion.
	I'm not convinced by the 90% trade figure with the UK. Can you provide any credible sources to back up this claim? Additionally, what specific impacts on Ireland's economy do you predict with a no-deal Brexit?
Comment	I think it's important to look at the bigger picture here. What about the impact on the global economy?
	I understand your point about the EU being 'terrified' of a no-deal Brexit, but what about the potential consequences for the UK?

Table 3.7: *Sample generated replies by Llama-2(13b) with different stances for a source tweet: "A No-deal Brexit will seriously damage Ireland; 90% of its trade is with the UK. It will cause mayhem in Northern France. Unemployment will reign in German car making cities. And it will be the unaccountable, unelected Eurocrats to blame.".*

for all the generations and per-stance averaged selfBLEU scores. A lower score indicates less similarity between each reply, reflecting greater diversity among the generations. As for the agreement with human judgements, we randomly sample three LLM generated replies per stance per source tweet (in total 138 replies), and manually annotate the stance of those replies towards the source tweets. The annotators do not have access to the silver stance labels during annotation. The percentage of the agreement reflects the quality of the silver labels. The results in Table 3.8 indicate that the generations exhibit a relatively high level of text diversity and a moderate to high label agreement. We also compare the silver labels

and human judgements for each stance as shown in Figure 3.3. The agreement over the *query* class is the highest, while the human annotators mostly disagree with the silver labelled *comments*. Over 30% of synthetic *comments* and *denies* are identified as *supports* by the human annotators. It is worth noting that an agreement of 50% for the highly important *deny* class is in line with agreement achieved by human annotators in the same task Li, Vasilakes, Zhao & Scarton (2025).

Metric	Agreement	selfBLEU2	selfBLEU3	selfBLEU4	selfBLEU2			
					S	D	Q	C
Value	0.64	0.0054	0.0026	0.0018	0.0048	0.0048	0.0062	0.0057

Table 3.8: *The percentage agreement between silver labels and human annotators and selfBLEU{n} scores (n denotes the n-gram level) for synthetic replies generated by Llama-2 (13b). Abbreviations: S, D, Q, C denote support, deny, query and comment respectively.*

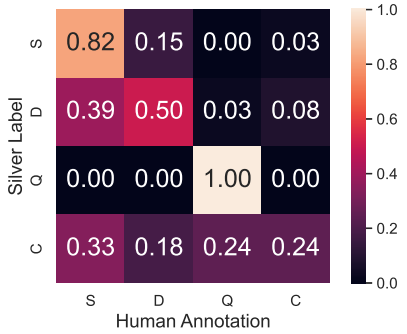


Figure 3.3: *The comparison between silver labels and human annotations for each stance. Each entry ij on row i column j denotes the proportion of synthetic replies with silver class i that is recognised as class j by human annotator.*

Furthermore, we manually inspect the generations and observe that the generated *queries* and *comments* are less natural or human-like, compared to the generated *supports* and *denies*. For example, as shown in Table 3.7, the generated *comments* typically follow a consistent structure or narrative, i.e. it tends to agree with the source tweet initially and then raise a question or doubt. The generated *queries* often start with "can you provide" or "I'd like to (e.g., know more)", which may be caused by the way the Llama-2 is instructed.

Self-training with Data Augmentation

Pseudo Labelling We summarise our self-training with data augmentation in Algorithm 1. We follow a standard self-training framework (Yarowsky 1995) based on iteratively generating pseudo labels D_{Tl} for unlabelled data in the target domain (rumours spread in Ireland) D_{Tu} using the source model M_S trained with the RumourEval data. The source model is then updated with the pseudo-labelled data in each iteration. Following Su et al. (2022), we utilise the model's predicted probability to measure model's confidence. The pseudo-labelled data whose probabilities are larger than a threshold τ are selected to optimise the source model.

Algorithm 1 The self-training algorithm coupled with synthetic data

Input:

M_S: Source model trained with RumourEval data
D_{Tu}: Unlabelled data in the target Irish domain
D_{Tl}: pseudo labelled data in the target Irish domain
D_{LLM}: Synthetic data generated by LLM
τ: Threshold for the selection of the pseudo labels
N_{iter}: The maximum number of iterations
S: The strategy of coupling with the synthetic data

$D_{Tl} \leftarrow \varnothing$
if $S = UpdateM_S$ **then**
    Fine tune M_S on D_{LLM}
end if
for $i \leftarrow 1, N_{iter}$ **do**
    if $D_{Tu} = \varnothing$ **then**
        Stop self-training
    else
        $D_{Tl_i} \leftarrow$ (for $d \in D_{Tu}$ if $M_S(d) > \tau$) ▷ $M_S(d)$ denotes the output probability of the model M_S
        if $D_{Tl_i} = \varnothing$ **then**
            Stop self-training
        else if $S = MixPseudo$ **then**
            $D_{Tl} \leftarrow D_{Tl} \cup D_{Tl_i}$
        else
            $D_{Tl} \leftarrow D_{Tl} \cup D_{Tl_i} \cup D_{LLM}$
        end if
    end if
    Fine tune M_S on D_{Tl}
end for

The whole process stops when there are no pseudo-labelled data satisfying this requirement or no more unlabelled data on the target domain.

As indicated in Algorithm 1 by strategy S, we consider the following two variants of the above self-training framework to integrate the silver-labelled synthetic data D_{LLM}.

- $S = UpdateM_S$. We continue fine-tuning the source model M_S that was trained with RumourEval data using the silver-labelled synthetic replies before the self-training process. We hypothesise that this model may be a better starting point for self-training, since it has been exposed to lexicons and knowledge in the target domain.
- $S = MixPseudo$. During self-training, we update the source model M_S with a mixture of the pseudo-labelled data D_{Tl_i} and synthetic data D_{LLM} in each iteration.

Additional Regularisation To further prevent the model from blindly trusting the false pseudo labels during self-training, we use a regularisation term in the loss function to encourage class diversification (Chen et al. 2022). The loss function for the self-training process can then be written as follows, where L^{ce} and L^{re} represent the cross-entropy loss and regularisation term on the predicted probability $p(x)$, respectively.

$$\begin{aligned} L^{Total} &= L^{ce} + L^{re} \\ &= -\mathbb{E}_{x \in D_{Tl}} \Sigma y \log p_y(x) + \mathbb{E}_{x \in D_{Tl}} \Sigma p(x) \log p(x) \end{aligned} \tag{3.3}$$

3.5.2 Experimental Setup

Source Model The source model, or any models fine-tuned with the RumourEval training set in this set of experiments, are the same BERTweet-large base model (Nguyen et al. 2020), as it achieves higher $wF2$ score on RumourEval validation set than Roberta-large model (Liu et al. 2019). We present its performance over `ISLES` in Table 3.9 (denoted as `FT[RumourEval]`).

Benchmarked Methods The proposed two self-training methods (`Self-training`[`S` = `UpdateM`$_S$] and `Self-training`[`S` = `MixPseudo`]) are compared to the following approaches and ablations:

- **Fine-tuning with the synthetic target domain data**. We consider fine-tuning a vanilla BERTweet (`FT[synthetic]`)) or a BERTweet pre-trained with RumourEval dataset (`FT[RumourEval+synthetic]`).
- **Fine-tuning with pseudo labels predicted by LLM**. We use Llama-2 (13b) to directly generate pseudo labels for the unlabelled target data, which are then used to fine-tune a vanilla BERTweet (`FT[Llama_pseudo]`) or a BERTweet pre-trained with RumourEval dataset (`FT[RumourEval+Llama_pseudo]`).
- **Self-training without data augmentation**. We use the source model trained with the RumourEval data to facilitate the self-training with unlabelled target data, without any data augmentation (`Self-training[w/o synthetic]`).

Unlabelled rumours spread in Ireland The full unlabelled data contains conversations by 12 source tweets of rumours spread in Ireland. To simulate a relatively low resource scenario, we do not utilise the full unlabelled data. Instead, 3,000 replies are randomly sampled as the unlabelled target data in our experiments. We further discuss the impact of the number of unlabelled target data in section 3.5.3.

Hyperparameters In practice, unsupervised adaptation lacks labelled data for hyperparameter tuning, therefore, following previous work (Su et al. 2022, Laparra et al. 2021), all of hyperparameters are set as the same as the source model without any hyperparameter tuning. This setting also illustrates the merit of hyperparameter insensitivity in our approach.

As for the confidence thresholding in self-training, the source model generates highly imbalanced distributed pseudo labels. We search the optimal threshold τ from $[0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$ and choose the highest value that ensures the minority class in the selected pseudo labels has at least three data points. Therefore, the threshold is set to 0.6 when no synthetic data is used, and the distribution of the pseudo labels for *support*, *deny*, *query* and *comment* classes is: $4 : 142 : 40 : 2360$. When synthetic data is added in the self-training process, we set the threshold to 0.9, following Su et al. (2022).

3.5.3 Results

The models' performance are presented in Table 3.9. Overall, integrating synthetically generated data to the source model before (`FT[RumourEval+synthetic]`) or during the self-training process ([$\texttt{S} = \texttt{UpdateM}_\texttt{S}$] and [$\texttt{S} = \texttt{MixPseudo}$]) yields better results than that without any data augmentation in self-training (`[w/o synthetic]`). The self-training method updated with synthetic data ([$\texttt{S} = \texttt{UpdateM}_\texttt{S}$]) achieves the best result, increasing the $wF2$ score with nearly 35 points (from 0.2519 to 0.5719), compared with the source model without any adaptation (`FT[RumourEval]`).

Method	$mF1$	GMR	$wF2$
`FT[RumourEval]`	0.3682	0.2914	0.2519
`FT[synthetic]`	0.2392	0.1464	0.2365
`FT[RumourEval+synthetic]`	0.5236	0.5216	0.5614
`FT[Llama_pseudo]`	0.4211	0.4721	0.4771
`FT[RumourEval+Llama_pseudo]`	0.4243	0.4716	0.4947
`Self-training`			
`[w/o synthetic]`	0.3960	0.4243	0.3733
[$\texttt{S} = \texttt{UpdateM}_\texttt{S}$]	**0.5320**	**0.5701**	**0.5719**
[$\texttt{S} = \texttt{MixPseudo}$]	0.4879	0.5100	0.4843

Table 3.9: *Performance of the source and adapted models evaluated on* `ISLES`*. The best performance is in bold, with statistically significant difference (t-test, p value* < 0.05*)*

Another important observation is that the synthetically generated replies by Llama-2 (13b) in the target domain may be more useful than the class predictions by Llama-2 (13b). Fine-

tuning the RumourEval-trained source model with synthetic data(`FT[RumourEval+synthetic]`) results in a better performance ($wF2 = 0.5614$) than the same source model fine-tuned with with pseudo labels generated by Llama-2 (`FT[RumourEval+Llama_pseudo]`, $wF2 = 0.4974$), even though the number of LLM pseudo-labelled instances (i.e., 3,000 tweets) is significantly larger than that of the synthetic data (i.e., 480 tweets). In fact, `FT[RumourEval+synthetic]` achieves the second best results on `ISLES`. Nevertheless, RumourEval data is still essential in this scenario: fine-tuning the vanilla BERTweet model with the synthetic data alone (`FT[synthetic]`) shows the worst performance (0.2365), even lower than the baseline source model (0.2519), illustrating that the model fine-tuned with the related source domain data serve as a suitable initialisation when only a small-sized synthetic target domain data is available.

Error Analysis `Self-training[w/o synthetic]` (Figure 3.4b) follows the same trend as the source model (Figure 3.4a): mis-classification of the *support* and *deny* classes into the *comment* class. Integrating the synthetic replies during the self-training process ([S = `MixPseudo`]) significantly enhance the performance for *support* and *deny* classes, but at the cost he *query* class' performance (Figure 3.4c), likely due to the introduction of the synthetic *queries* with relatively low naturalness.

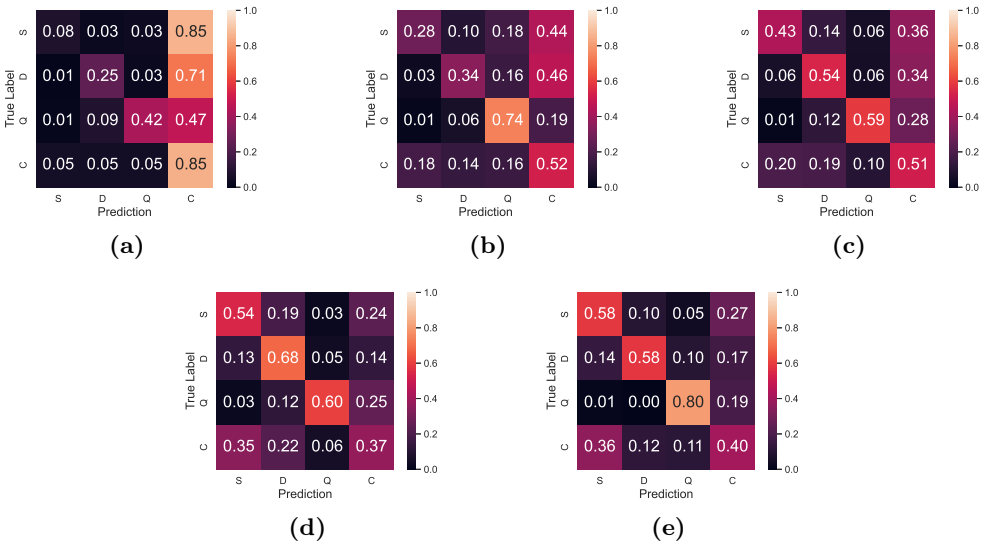
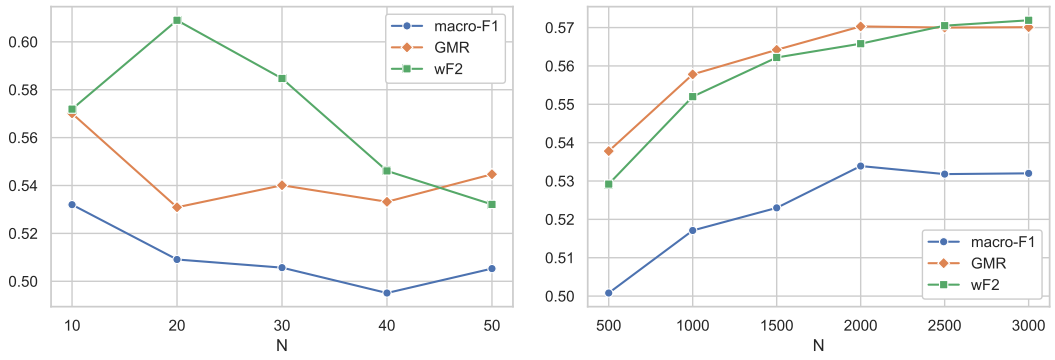


Figure 3.4: *Confusion matrices of different models assessed on* `ISLES`*. (a) Baseline source model (*`FT[RumourEval]`*); (b)* `Self-training[w/o synthetic]`*; (c)* `Self-training`[S = `MixPseudo`]*; (d)* `FT[RumourEval+synthetic]`*; (e)* `Self-training`([S = `UpdateM`$_S$]*).*

Similarly, the approaches that involve fine-tuning with synthetic data are also influenced by the quality of the generated replies. After fine-tuning the RumourEval-trained source model with the synthetic replies (`FT[RumourEval+synthetic]`, Figure 3.4d), the performance over the *support* and *deny* classes is highly improved while the *comment* class that contains less human-like generations is significantly harmed when compared to the source model performance (Figure 3.4a). The unnatural synthetic *queries* could still benefit the learning process

because they feature certain real-world characteristics, such as question marks and interrogative pronouns. Finally, `Self-training`[$\mathtt{S} = \mathtt{UpdateM_S}$] improves the performance for both *query* and *comment* classes (Figure 3.4e), while maintaining a high performance for both *support* and *deny*. Another error produced by models with synthetic data is the mis-classification of the *comment* class into *support*, which aligns with our analysis of the silver labels whose *comments* are often recognised as *supports* by human annotators (see Figure 3.3).

Impact of the amount of synthetic data To analyse whether increasing the number of synthetically generated replies could lead to further improvement, we generate 20, 30, 40 or 50 replies per stance per source tweet with Llama-2 (13b) and experiment with our best-performing self-training method ([$\mathtt{S} = \mathtt{UpdateM_S}$]). As shown in Figure 3.5a, increasing the number of synthetic data to more than 20 replies lead to worse $wF2$ scores. Training with more synthetic generations tends to also decrease the GMR and $mF1$ scores. This behaviour can be explained by the mis-classification of *comments* into *supports* in the synthetic data, as we discussed in the error analysis. This tendency of mis-classification gets more severe when adding more synthetic data, leading to higher recall scores on the *support* and *deny* classes (higher $wF2$ score at 20 examples), while lower GMR score which penalises the model with bad recall on the *comment* class.



(a) **(b)**

Figure 3.5: *The impact of the amount of synthetic data and unlabelled data on the $wF2$ score for the best-performing self-training model ([$\mathtt{S} = \mathtt{UpdateM_S}$]). (a) The impact of the amount of synthetic data (N: the number of replies generated per stance per source tweet); (b) The impact of the amount of unlabelled target domain data (N: the number of unlabelled replies).*

Impact of the amount of unlabelled real-world target domain data To analyse whether our best-performing method is still effective with even less unlabelled real-world target domain data, we randomly sample 2,500, 2,000, 1,500, 1,000, and 500 tweets from the 3,000 unlabelled tweets we used in the previous experiments and re-run the `Self-training` [$\mathtt{S} = \mathtt{UpdateM_S}$] model. As shown in Figure 3.5b, decreasing the amount of unlabelled target domain data harms the model performance. When the number of unlabelled tweets is lower than 1,500, the $wF2$ score is worse than the source model before self-training (i.e. `FT[RumourEval+synthetic]`, $wF2 = 0.5614$), illustrating that self-training may not be bene-

ficial without enough real-world examples. In contrast, self-training always shows high GMR scores even with 500 unlabelled replies, which can be explained by real-world data providing information for improving the classification of *comments* (not well simulated on synthetic data). Nevertheless, acquiring over 1,500 unlabelled data for self-training is significantly cheaper than annotating more real-world for fine-tuning, being a suitable approach for rumour stance classification.

3.6 Conclusion

This paper explores domain adaption for rumour stance classification. We first examine the generalisability of current models by introducing a novel test set (`ISLES`) for this task, which differs significantly from the widely used RumourEval dataset regarding topic and lexicon coverage and stance distribution. For the supervised models trained with the RumourEval data, our evaluation on `ISLES` shows substantial performance drop compared to their performances on the RumourEval test data. More importantly, the performance rankings of these models on the RumourEval test set do not align with their rankings on ISLES, underscoring the limitation of current benchmarks in measuring true model generalisability for rumour stance classification. Although zero-shot in-context learning with LLMs lags behind most supervised models when both training and testing are performed on the RumourEval data, it exhibits stronger generalisability and better performance when evaluating on the `ISLES` dataset, highlighting its potential under significant domain shift.

Next, we investigate a novel source-free unsupervised adaptation setting for rumour stance classification. We propose a new self-training framework incorporating with synthetic data generated by LLMs. Our approach achieves an improvement of nearly 127% over the baseline performance prior to adaptation, using only synthetic generated and real-world unlabelled target-domain data. Our analysis shows that fine-tuning with synthetic texts generated by LLMs could be more effective than using real-world data pseudo-labelled by LLMs. Furthermore, we highlight the crucial role of synthetic data quality, and demonstrate the importance of incorporating unlabelled real-world data to bridge the gap between synthetic and real-world data differences.

Limitations and Future Work

In this work, we only focus on English rumour stance classification. More than 89.70% of the pre-training data of Llama-2 is English, and its performance on generating synthetic replies in languages other than English is under-explored. However, our framework could be easily adapted to handle non-English languages with the help of LLMs that are specialised or trained in the respective non-English languages. Future work could also explore the cross-lingual rumour stance classification under our framework, treating the source and target languages as different domains.

Another limitation of our work is that we assume the source and target data share exactly the same stance categories (i.e. closed-set adaptation). However, in real-world scenarios, this assumption may not hold true. For example, in situations of extremely low data resources, the collected unlabelled target data may not encompass texts representing all of the stance

categories used in the source data. Therefore, future research can investigate the open-set source-free unsupervised adaptation for rumour stance classification.

Finally, due to the lack of manual annotations for the unlabelled target domain data used during adaptation, we do not quantify or estimate the quality of the pseudo labels used during self-training, which can provide a useful insight for approaches to increase the quality of the pseudo labels.

Acknowledgments

This work was supported by the European Union under action number 2020-EU-IA-0282 and agreement number INEA/CEF/ICT/A2020/2381686 (EDMO Ireland[15]) and the UK's innovation agency (InnovateUK) grant number 10039039 (approved under the Horizon Europe Programme as VIGILANT[16], EU grant agreement number 101073921). Yue Li is supported by a Sheffield–China Scholarships Council PhD Scholarship. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

3.7 Appendix

3.7.1 Dataset Supplementary Information

Our collected rumours, their veracity and the number of reply tweets are listed in Table 3.10. Each rumour has one corresponding source tweet, except for "*Donald Tusk said there was a special place in hell for Brexiteers*" and "*Dynamo Kyiv players in military uniforms prepared to fight Russian troops in Ukraine*", which have two related source tweets.

The vocabulary overlap between RumourEval dataset and each group is shown in Figure 3.6. Each entry ij denotes the proportion of words in the group/dataset in row i that are also present in the group/dataset in column j. We sample our test set from each topic group based on the ratio of values in the last column.

3.7.2 Annotation Guideline and Interface

Annotation Guideline

Introduction The figure 3.7 illustrates a conversation thread, consisting of a source tweet (from user 1) related to a rumour story and its direct (e.g., the tweets from user 2 and 3) and indirect replies (e.g., the tweet from user 4). Each indirect reply tweet has a parent tweet to which it directly replies (e.g., the tweet from user 3 is the parent tweet of that from user 4).

The annotators need to infer the stance of the replies (i.e. target tweets) towards the source tweet. During annotation, the thread is split into triples (source tweet, parent tweet, target tweet) for indirect replies and tuples (source tweet, target tweet) for direct replies.

Stance Categories

[15] `https://edmohub.ie`

[16] `https://www.vigilantproject.eu`

Rumour	Verdict	# of replies	Topic
Covid-19 vaccination figures being reported by the media were false as they included missed appointments.	False	113	COVID-19
Bulgarian workers queued outside a Lidl supermarket to pick fruit in Dublin during COVID-19 pandemic.	False	298	COVID-19
Restaurants were forced to close in 3 hours after Dublin went into Level 3 restrictions for COVID-19.	False	290	COVID-19
Donald Tusk said there was a special place in hell for Brexiteers.	False	885	Brexit
Irish Taoiseach threatened to ban British planes from Irish skies.	False	490	Brexit
Northern Ireland drivers will have to display GB stickers on their vehicles if driving in Ireland after Brexit.	Unproven	124	Brexit
90% of Ireland's trade was with the UK.	False	7701	Brexit
The Irish government has cut the mental health budget by €20 million.	False	46	Irish politics
The Irish government has quadrupled social housing construction in two years.	Half True	56	Irish politics
Children aged 4 and under would be taught about masturbation if the sex education bill becomes law in Ireland.	False	154	Irish politics
The Irish government will spend more on HAP than on new social housing.	Mostly True	61	Irish politics
Dynamo Kyiv players in military uniforms prepared to fight Russian troops in Ukraine.	False	114	Ukraine war
Ukrainian refugees and asylum seekers were segregated by race in a Dublin processing centre.	False	20	Ukraine war
Congresswoman Ilhan Omar was married to her brother.	False	696	Other

Table 3.10: *Collected rumours, corresponding verdict, topic group and the number of reply tweets before data filtering and sampling*

- **Confirming:** The author of the response supports the source tweet. For example[17]:

  Source tweet: AFP reports there are 2 dead, and 5 hostages being held in the Kosher store in Eastern Paris; separate incident to charlie hebdo shooters.

  Target tweet: @USER Yes. It was confirmed officially today. [Confirming]

- **Rejecting:** The author of the response disagrees with source tweet. For example:

  Source Tweet: We understand there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney.. ISIS flags remain on display #7News

  Target Tweet: @USER several of your tweets are either wrong or misleading. Please hold yourself to higher journalism standards! [Rejecting]

  Note that sarcastically or humorously supporting the source tweet should be considered as rejecting, while sarcastically or humorously denying source tweet should be considered as confirming.

[17]We provide annotators with examples in RumourEval 2019 dataset.

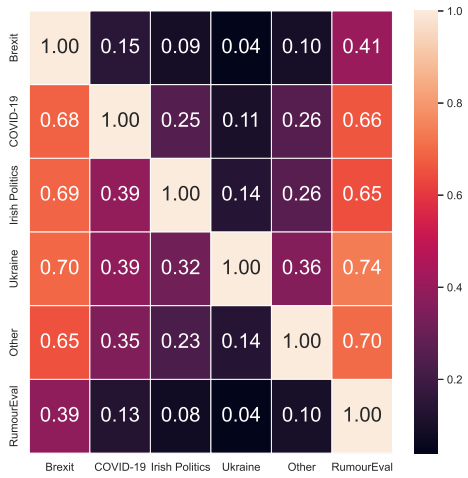


Figure 3.6: *The vocabulary overlap between RumourEval dataset and each group*

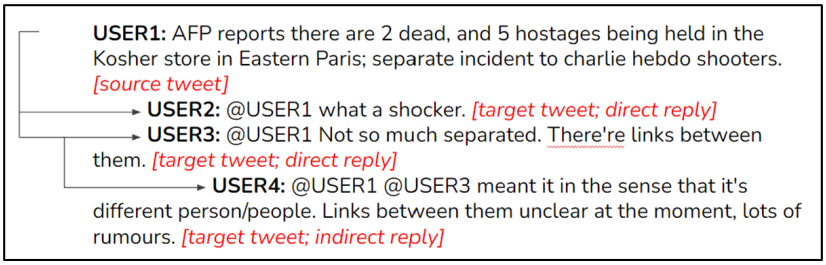


Figure 3.7: *Example of a Twitter conversation thread from RumourEval 2019 dataset.*

- **Questioning:** The author of the response asks for additional evidence in relation to the source tweet. For example:

  Source tweet: We understand there are two gunmen and up to a dozen hostages inside the cafe under siege at Sydney.. ISIS flags remain on display #7News

  Target tweet: @USER @USER sorry - how do you know it's an ISIS flag? Can you actually confirm that? [Questioning]

  Note that the appearance of question marks does NOT necessarily indicate the questioning stance. For example, rhetorical questions should not be coded as questioning, since the author does not expect an answer.

  Source tweet: Clinton camp delays Weather Channel ad buy after backlash

  Target tweet: @USER Sad it had to get to that. What was she thinking in the 1st place? [Commenting]

- **Commenting:** The author of the response makes their own comment without a clear stance towards the source tweet. For example:

Source tweet: 11 confirmed dead, Francois Hollande to visit scene of attack - latest from Paris.

Target tweet: Hi @USER this is the photo of our journalist @USER please credit him at least [Commenting]

Annotation Interface

Our annotation interface is shown in Figure 3.8. Note that "target" refers to the reply tweet to be annotated in Figure 3.8 while it refers to the target of the stance (i.e., source tweet) in Chapter 2.

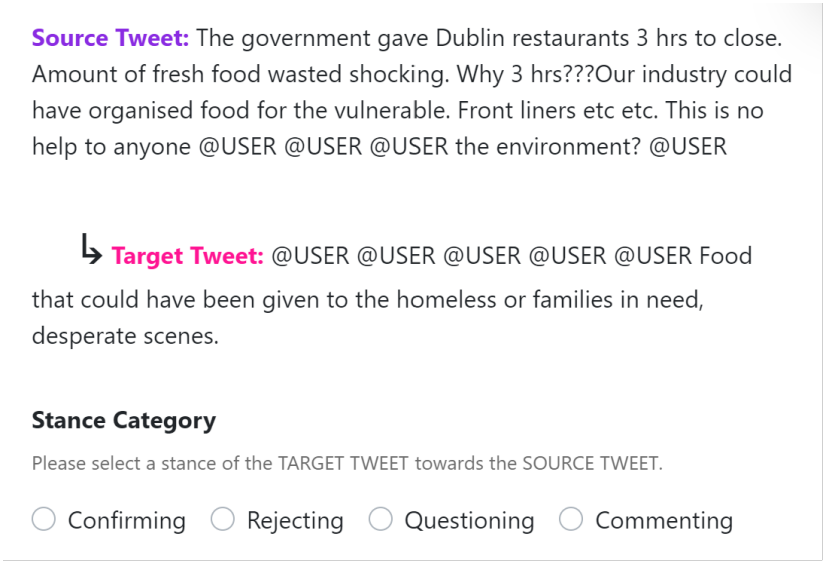


Figure 3.8: *Annotation interface for Irish rumour stance classification.*

3.7.3 Model Details

We present the model parameters that deviate from the previous work, or no previous work to refer to.

BERTweet and RoBERTa We use the *bertweet-large*[18] and *RoBERTa-large*[19]. We adopt class weighted loss function. The class weights are computed according to the class distribution of the RumourEval 2019 training data. We use grid search for hyperparameter tuning and the optimal hyperparameters are determined based on the $wF2$ score on the validation set. For each model, we repeat the fine-tuning process for five times with different random seeds.

As for fine-tuning, we employ the transformers library (Wolf et al. 2020) and adopt AdamW (Loshchilov & Hutter 2019). We search the batch size from [16, 32], the learning rate from [1e-5,3e-5,5e-5,7e-5,1e-4], and the number of epochs from [3, 4, 5, 6, 7, 8, 9, 10]. As for

[18] `https://huggingface.co/vinai/bertweet-large`
[19] `https://huggingface.co/FacebookAI/RoBERTa-large`

prompt-tuning, we use the implementation of Liu, Ji, Fu, Tam, Du, Yang & Tang (2022) [20]. We search the prompt length from [10, 20, 30], the batch size from [16, 32], and the learning rate from [1e-2,2e-2,3e-2]. We set the maximum number of epochs as 50, and adopt early-stopping strategy for prompt-tuning.

Hierarchical BERT We adopt the implementation of Yu et al. (2020) for the single task model[21] for rumour stance detection and re-train it with the RumourEval 2019 training set. We use the hyperparameters suggested by the authors. However, due to memory limitations, we reduce the batch size from 2 to 1 and tune the learning rate from [1e-5,3e-5,5e-5,7e-5,1e-4]. We repeat the training process for five times with different random seeds.

Llama and Mistral For zero-shot in-context learning, We test the performance of the following 12 prompts for Llama and Mistral models, and choose the optimal one based on its performance over the RumourEval validation set. Basically, we try to phrase the instructions in various ways and try different words to denote stance classes, inspired by the stance names used in other stance classification datasets (e.g., in favour, against etc) (Hardalov et al. 2021).

As for the post-processing of the model generation to extract the stance class, we adopt regular expression to match the stance names used in the prompt. If there are no matched stance names (very few cases), we assume it belongs to the comment class, since it is the majority class of the RumourEval dataset.

1. What is the stance of tweet B regarding the tweet A? Choose the stance category from: 1. Support: tweet B agrees with tweet A; 2. Deny: tweet B disagrees with tweet A; 3. Query: tweet B asks for additional evidence related to tweet A; 4. Comment: tweet B makes their own comment without a clear stance towards tweet A. Tweet A: {text1}. Tweet B: {text2}.

2. What is the stance of tweet B regarding the tweet A? Options: [<Support>: tweet B agrees with tweet A; <Deny>: tweet B disagrees with tweet A; <Query>: tweet B asks for additional evidence related to tweet A; <Comment>: tweet B makes their own comment without a clear stance towards tweet A]. Tweet A: {text1}. Tweet B: {text2}.

3. Given a source tweet and its reply, detect the stance that the reply has towards the source tweet. There are four options: <support>, <deny>, <query> and <comment>. If the reply supports the source tweet, answer with <support>; if the reply opposes the source tweet, answer with <deny>; if the reply asks for additional evidence in relation to the source tweet, answer with <query>; if the reply makes their own comment without a clear stance, answer with <comment>. Now complete the following example. Source tweet: {text1}. Reply: {text2}.

4. What is the stance of tweet B regarding the tweet A? Choose the stance category from: 1. Supporting: tweet B agrees with tweet A; 2. Refuting: tweet B disagrees with tweet A; 3. Questioning: tweet B asks for additional evidence related to tweet A; 4. Commenting: tweet B makes their own comment without a clear stance towards tweet A. Tweet A: {text1}. Tweet B: {text2}.

[20] https://github.com/THUDM/P-tuning-v2

[21] https://github.com/jefferyYu/DualHierarchicalTransformer

5. What is the stance of tweet B regarding the tweet A? Options: [<Supporting>: tweet B agrees with tweet A; <Refuting>: tweet B disagrees with tweet A; <Questioning>: tweet B asks for additional evidence related to tweet A; <Commenting>: tweet B makes their own comment without a clear stance towards tweet A]. Tweet A: {text1}. Tweet B: {text2}.

6. Given a source tweet and its reply, detect the stance that the reply has towards the source tweet. There are four options: <supporting>, <refuting>, <questioning> and <commenting>. If the reply supports the source tweet, answer with <supporting>; if the reply opposes the source tweet, answer with <refuting>; if the reply asks for additional evidence in relation to the source tweet, answer with <questioning>; if the reply makes their own comment without a clear stance, answer with <commenting>. Now complete the following example. Source tweet: {text1}. Reply: {text2}.

7. What is the stance of tweet B regarding the tweet A? Choose the stance category from: 1. Agree: tweet B agrees with tweet A; 2. Disagree: tweet B disagrees with tweet A; 3. Query: tweet B asks for additional evidence related to tweet A; 4. Comment: tweet B makes their own comment without a clear stance towards tweet A. Tweet A: {text1}. Tweet B: {text2}.

8. What is the stance of tweet B regarding the tweet A? Options: [<Agree>: tweet B agrees with tweet A; <Disagree>: tweet B disagrees with tweet A; <Query>: tweet B asks for additional evidence related to tweet A; <Comment>: tweet B makes their own comment without a clear stance towards tweet A]. Tweet A: {text1}. Tweet B: {text2}.

9. Given a source tweet and its reply, detect the stance that the reply has towards the source tweet. There are four options: <agree>, <disagree>, <query> and <comment>. If the reply supports the source tweet, answer with <agree>; if the reply opposes the source tweet, answer with <disagree>; if the reply asks for additional evidence in relation to the source tweet, answer with <query>; if the reply makes their own comment without a clear stance, answer with <comment>. Now complete the following example. Source tweet: {text1}. Reply: {text2}.

10. What is the stance of tweet B regarding the tweet A? Choose the stance category from: 1.In favour: tweet B agrees with tweet A; 2. Against: tweet B disagrees with tweet A; 3. Query: tweet B asks for additional evidence related to tweet A; 4. Comment: tweet B makes their own comment without a clear stance towards tweet A. Tweet A: {text1}. Tweet B: {text2}.

11. What is the stance of tweet B regarding the tweet A? Options: [<In favour>: tweet B agrees with tweet A; <Against>: tweet B disagrees with tweet A; <Query>: tweet B asks for additional evidence related to tweet A; <Comment>: tweet B makes their own comment without a clear stance towards tweet A]. Tweet A: {text1}. Tweet B: {text2}.

12. Given a source tweet and its reply, detect the stance that the reply has towards the source tweet. There are four options: <in favour>, <against>, <query> and <comment>. If the reply supports the source tweet, answer with <in favour>; if the reply opposes the source tweet, answer with <against>; if the reply asks for additional evidence in

relation to the source tweet, answer with <query>; if the reply makes their own comment without a clear stance, answer with <comment>. Now complete the following example. Source tweet: {text1}. Reply: {text2}.

Chapter 4

Publication III: Label Set Optimization via Activation Distribution Kurtosis for Zero-Shot Classification with Generative Models

Overview Publications I and II investigate the generalisation and adaptation of supervised small/medium-sized pre-trained language models (PLMs), which require substantial amount of human-labelled or pseudo-labelled training data. However, data collection and annotation are both time-consuming and labour-intensive. Since LLMs have demonstrated superior ICL performance compared to supervised PLMs across various NLP tasks, the publication presented in this chapter explores the generalisation ability of LLMs for rumour stance classification in a zero-shot ICL setting, where no task-specific training data is required. We tackle the well-known label adaptation problem in stance classification and propose a novel, data-efficient method for identifying optimal stance label names within prompts. Our approach effectively improves generalisation performance in zero-shot ICL for rumour stance classification.

Label Set Optimization via Activation Distribution Kurtosis for Zero-Shot Classification with Generative Models

Yue Li, Zhixue Zhao, Carolina Scarton

School of Computer Science, University of Sheffield, UK

Abstract

In-context learning (ICL) performance is highly sensitive to prompt design, yet the impact of class label options (e.g. lexicon or order) in zero-shot classification remains underexplored. This study proposes LOADS (Label set Optimization via Activation Distribution kurtosiS), a post-hoc method for selecting optimal label sets in zero-shot ICL with large language models (LLMs). LOADS is built upon the observations in our empirical analysis, the first to systematically examine how label option design (i.e., lexical choice, order,

and elaboration) impacts classification performance. This analysis shows that the lexical choice of the labels in the prompt (such as *agree* vs. *support* in stance classification) plays an important role in both model performance and model's sensitivity to the label order. A further investigation demonstrates that optimal label words tend to activate fewer outlier neurons in LLMs' feed-forward networks. LOADS then leverages kurtosis to measure the neuron activation distribution for label selection, requiring only a single forward pass without gradient propagation or labelled data. The LOADS-selected label words consistently demonstrate effectiveness for zero-shot ICL across classification tasks, datasets, models and languages, achieving maximum performance gain from 0.54 to 0.76 compared to the conventional approach of using original dataset label words.

4.1 Introduction

Generative large language models (LLMs) are increasingly used for classification tasks via zero-shot in-context learning (ICL), where models are prompted to select an option from a pre-defined set of labels (Wang et al. 2022, Antypas et al. 2023*a*, Gonen et al. 2023, Mu et al. 2024). While some classification tasks employ a relatively fixed set of lexicons to represent class labels, such as sentiment analysis (*positive* and *negative*) and textual entailment (*entailment* and *contradiction*), other tasks may present more ambiguous choices in lexical selection. Stance classification, for instance, uses diverse pairs of antonyms to represent positive and negative stances across different datasets, e.g. *agree-disagree* vs. *favor-against*. As a result, when crafting prompts for these classification tasks, practitioners face decisions regarding label options in the prompt, such as lexical selection and ordering.

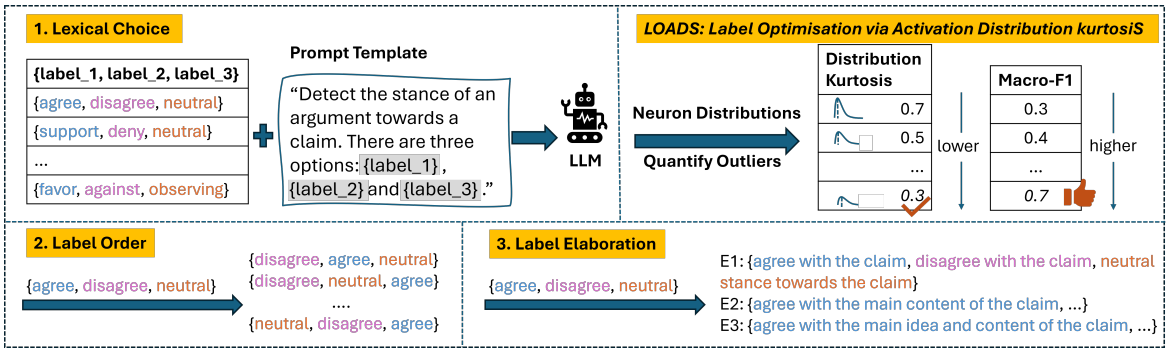


Figure 4.1: *Illustration of the three aspects (i.e., lexical choice, label order and label elaboration) for designing the label option in the prompt in zero-shot ICL for classification, and our* LOADS *to post-hoc select the optimal label set (top half figure).*

Despite studies suggesting the sensitivity of ICL to prompt design (Lu et al. 2022, Yoo et al. 2022, Wei et al. 2024, Mao et al. 2024, Liu, Lin, Hewitt, Paranjape, Bevilacqua, Petroni & Liang 2024, Zhang et al. 2022, Liu, Shen, Zhang, Dolan, Carin & Chen 2022, Peng et al. 2024, Gonen et al. 2023, Mu et al. 2024), this subtle yet critical consideration of label options in prompt for zero-shot ICL has received limited attention. To fill in this research gap, we explore the impact of three types of label variants (i.e., lexical choice, label order and elaborations)

in zero-shot ICL with both encoder-decoder and decoder-only LLMs. We mainly ground our research on stance classification, a task where label adaptation is a known problem due to various label inventories in different studies (Hardalov et al. 2021). We demonstrate that the lexical choice of the label options significantly impacts model performance. The model's sensitivity to the label order also depends on the lexical choice, while elaborating on task-related information (e.g. *agree with the claim* elaborating *agree*) has minimum effect.

Inspired by recent studies on neuron analysis (Kuzmin et al. 2023, Voita et al. 2024, Stolfo et al. 2024, Kurz et al. 2024), we further investigate the neurons in the feed-forward network (FFN) in the decoder of the LLMs. We empirically show that prompts with optimal label sets activate fewer outlier neurons. Consequently, we propose a new method, **L**abel set **O**ptimization via **A**ctivation **D**istribution kurtosi**S** (LOADS), to select optimal label sets for a given classification dataset in zero-shot ICL. LOADS could stably and effectively work with only 100 unlabelled samples of the validation dataset, also demonstrating transferability across datasets and languages. Our contributions are summarized as follows:

- The first benchmark on **how variants of label options in prompts affect zero-shot ICL models' performance** for classification tasks. We provide useful recommendations on label designing to practitioners working on zero-shot classification with LLMs [1].
- The empirical demonstration that **zero-shot ICL performance negatively correlates with the number of outlier neurons in FFN** when varying the lexical choices for label options. The correlation holds true across diverse English stance classification datasets and topic classification datasets with different models.
- A **novel and efficient post-hoc method (LOADS) for label selection in zero-shot ICL**. Compared with common strategies in practice, our approach demonstrates statistically significant performance improvements across model types, model sizes and languages with only 100 unlabeled data samples. Our analysis also suggests that the LOADS-selected label set is potentially transferable across similar datasets for a specific LLM, further alleviating the cost to collect samples for a target new dataset.

We present our experimental setups, results and discussions on the impact of label options in zero-shot ICL in Section 4.3. Then, we describe our neuron analysis of the lexical choice in label options in Section 4.4 and our proposed method LOADS for selecting optimal label sets in Section 4.5.

4.2 Related Work

ICL Performance Few-shot ICL mainly focuses on cases where LLMs are directly prompted with N demonstration examples. The studies highlight the substantial impact of example ordering (Lu et al. 2022), formatting (Yoo et al. 2022, Wei et al. 2024, Mao et al. 2024, Liu, Lin, Hewitt, Paranjape, Bevilacqua, Petroni & Liang 2024), and examples selection (Zhang et al. 2022, Liu, Shen, Zhang, Dolan, Carin & Chen 2022, Peng et al. 2024). The parallel lines

[1] Code and resources: `https://github.com/YLi999/Stance_LOADS`.

of work focus on improving few-shot ICL via the optimal selection or arrangement of examples (Liu, Shen, Zhang, Dolan, Carin & Chen 2022, Rubin et al. 2022, Lu et al. 2022, Zhang et al. 2022, Liu, Liu, Shi, Cheng & Lu 2024, Xu et al. 2024), re-weighting examples (Yang et al. 2023), automatic reformat or generation of demonstration representations (Kim et al. 2022, Liu et al. 2023), and introduction of intermediate reasoning steps (Wei, Wang, Schuurmans, Bosma, Xia, Chi, Le, Zhou et al. 2022, Zhang, Zhang, Li & Smola 2023). However, the impact of lexical choices for label names in classification received little attention, with the only closely related work suggesting that LLMs are likely to confuse classes which share similar key vectors in the attention modules (Wang et al. 2023).

For zero-shot ICL, Mu et al. (2024) demonstrate the effect of using synonyms for class options, but they neither consider the order of the label options nor propose an effective strategy to choose the label names. Gonen et al. (2023) empirically show that perplexity could be an effective indicator for prompt selection, but they do not account for class options. Notably, behavioral differences between few-shot and zero-shot ICL have been frequently observed, suggesting that findings from few-shot ICL do not necessarily hold in the zero-shot context (Lin & Lee 2024).

Prompt-Tuning and Verbalizer Prompt-tuning aims to automatically find or generate an optimal discrete prompt (e.g., through gradient-based search (Shin et al. 2020, Shi et al. 2023) or fine-tuning (Gao et al. 2021, Le Scao & Rush 2021, Deng et al. 2022)) or by training continuous prompt (Lester et al. 2021, Liu, Ji, Fu, Tam, Du, Yang & Tang 2022). Verbalizer can be taken as a mapping function that links discrete class labels to corresponding tokens or phrases in a model's vocabulary. A range of methods developed to build the verbalizer, including manually created verbalizer (Schick & Schütze 2021), search-based verbalizer that identifies label words automatically from the dataset (Gao et al. 2021, Shin et al. 2020), and soft verbalizers that uses continuous embeddings obtained through fine-tuning (Hambardzumyan et al. 2021, Cui et al. 2022). Prompt-tuning often does not focus on label set selection for zero-shot ICL, and the verbalizer introduces additional components to the decoding of the generative models, distinguishing it fundamentally from our work.

4.3 Prompting with Varied Label Options for Zero-shot Classification

In zero-shot ICL for classification, a common approach is to provide a set of class label options in the prompt to instruct the LLMs to choose one of the options as the classification prediction. Although the label option is a subtle component in the prompt, we are interested in whether it has a significant impact on model performance.

Specifically, we explore three types of variants around label options in the prompt: (1) *lexical choice*; (2) *label order*; and (3) *label elaboration*. To accurately measure the impact of these factors, we only manipulate the label options within the same prompt template (Section 4.3.2). We show examples of the three variants in Figure 4.1.

4.3.1 Methodology

Lexical Choice We use single-word synonyms to represent class labels (e.g. *support* and *agree*), forming various label sets. For each dataset, we compare the zero-shot ICL performance when LLMs are prompted to select from different label sets, as illustrated in Figure 4.1. For this purpose, we design a pipeline to create a pool of label sets:

1. *Collect a seed set of label names.* We obtain this set by collecting the label names in the datasets we experiment with.

2. *Expand label sets with WordNet and LLMs.* We use WordNet (Fellbaum 1998) as a reliable source and Claude[2] as a supplementary source to obtain synonyms for label names in the seed set. For pairs of label names with semantically opposite meanings (such as "agree" and "disagree"), we also consider antonyms to avoid potential ambiguity and present clear contrast for the predicted models.

3. *Manual selection.* We manually filter out semantically unrelated or inappropriate label sets generated by Claude to mitigate the impact of noisy label names.

The label names are arranged in the sequence presented in their original study (see Table 4.1). We refer to this arrangement as the *default order*.

Label Order We consider every possible order of the single-word labels in the prompt and compare the model performance against that obtained with the default order. For binary datasets, there is only one alternative arrangement besides the default order, while N-way multi-class classification would yield $N! - 1$ alternative orders.

Label Elaboration We investigate whether transforming single-word labels (e.g., "agree") into more detailed phrases (e.g., "agree with the claim") has an impact on model performance. On the one hand, elaborating on task details with the label may provide the model with a stronger alignment signal between the label and the task, emphasizing what the label is referring to. On the other hand, it also increases the label length and may introduce noise in the prompt (Liu, Lin, Hewitt, Paranjape, Bevilacqua, Petroni & Liang 2024). Therefore, we design three levels of elaborations (shorted for *E1*, *E2* and *E3*) by progressively adding more task-related (and potentially redundant) information to the single-word label names, as shown in Figure 4.1.

4.3.2 Experimental Setups

Datasets We focus on the stance classification task due to its rich label inventories across various readily available datasets. Stance classification aims to identify the type of an expressed opinion (e.g., "agree" or "disagree") in a given piece of text towards a particular topic, claim, or entity. We consider four binary (*scd* (Hasan & Ng 2013), *perspectrum* (Chen et al. 2019), *snopes* (Hanselowski et al. 2019) and *ibmcs* (Bar-Haim et al. 2017)) and five multi-class datasets (*vast* (Allaway & McKeown 2020), *emergent* (Ferreira & Vlachos 2016), *semeval* (Mohammad et al. 2016), *rumoureval* (Gorrell et al. 2019*a*) and *arc* (Hanselowski et al. 2018))

[2] `https://claude.ai/new`

from existing English stance classification benchmarks (Schiller et al. 2021, Hardalov et al. 2021, Chen et al. 2023), as shown in Table 4.1. The nine datasets cover different domains, such as social media posts, news articles and online debates forums. We also experiment with topic classification in Section 4.4 and 4.5 to demonstrate the generalizability of our findings to other NLP tasks.

Dataset Name	Original Label Words	Optimal Label Words
scd	for, against	pro, con
perspectrum	support, undermine	validate, refute
snopes	agree, refute	affirm, refute
ibmcs	pro, con	endorse, deny
vast	pro, con, neutral	confirm, dispute, neither
emergent	for, against, observing	endorse, reject, neutral
semeval	favour, against, neither	accept, reject, neutral
rumoureval	support, deny, query, comment	confirm, reject, question, neutral
arc	agree, disagree, discuss, unrelated	affirm, refute, discuss, unrelated

Table 4.1: *Lists of the English stance classification datasets, their labels in original dataset, and the optimal labels with the highest zero-shot ICL performance on Flan-T5-xl as an example to justify our motivation on* LOADS.

Models We cover both encoder-decoder and decoder-only LLMs, and experiment with the prevalent open-sourced Flan-T5 (Chung et al. 2024), Llama 3 and Llama 3.1 (Dubey et al. 2024) model families as representatives for these two types of LLMs. We choose their moderate-sized instruction-tuned versions, Flan-T5-xl (3b), Llama -3-Instruct (8b) and Llama-3.1-Instruct (8b), as our primary models for investigation due to our hardware resources constraints and their decent zero-shot ICL performance (Aiyappa et al. 2024, Chung et al. 2024, Dubey et al. 2024). We conduct experiments with *Gemma-2-it (9b)* [3] and *Flan-T5-xxl* (13b) to further show the generalizability of LOADS in Section 4.5.

Prompt Template We refer to the prompt template used in the supervised fine-tuning of Flan-T5 and Llama (Wang et al. 2022, Chung et al. 2024, Dubey et al. 2024). We present the results with the following prompt template in this paper: *Given a [text1_name] and a [text2_name], detect the stance that the [text2_name] has towards the [text1_name]. There are {N} options: "{label$_0$}", "{label$_1$}, ... , and {label$_{N-1}$}". Now complete the following example. [text1_name]: {text1}. [text2_name]: {text2}.* We also test other templates and find the results are consistent (e.g., prompting with label explanations in Appendix 4.7.4).

Evaluation We adopt macro-$F1$ for model performance evaluation to align with prior studies (Schiller et al. 2021, Hardalov et al. 2021, Chen et al. 2023). We use $wF2$[4] to account for data imbalance in rumoureval (Scarton et al. 2020*a*).

[3] https://huggingface.co/google/gemma-2-9b-it

[4] $wF2$ gives different weights for each stance: *deny* = *support* = 0.40, *query* = 0.15 and *comment* = 0.05.

Implementation Details To ensure reproducibility, we use greedy search for decoding[5]. In more than 95% cases, LLMs exactly follow the instruction and output the stance name within the required stance options. Therefore, we directly use the model generation as the predicted label without post-processing or mapping. We run experiments on the validation set of each dataset. See Appendix 4.7.3 for details on the label sets experimented with. We exclude the topic/entities-based (such as stance towards Obama) stance classification datasets (i.e. scd, semeval and vast) in label elaboration experiments to avoid unnecessary ambiguity or bias during elaboration (e.g., the text to be classified may refer to anything from policy to personal behavior about Obama, and elaborating *agree* to *agree with the opinion of Obama* could lead to biased prediction).

4.3.3 Results and Analysis

We first discuss the impact of lexical choice, label order, and label elaboration on zero-shot ICL for classification. Then we provide suggestions for practitioners in zero-shot ICL for classification.

	F_{score}	perspectrum	ibmcs	snopes	scd	emergent	semeval	rumoureval	arc	vast
Llama 3 (8b)	*max*	0.869	0.834	0.770	0.759	0.740	0.688	0.659	0.453	0.382
	min	0.738	0.592	0.540	0.639	0.387	0.544	0.286	0.249	0.173
	Original	0.799	0.679	0.656	0.709	0.467	0.643	0.521	0.364	0.219
	avg±var	0.824±0.001	0.756±0.003	0.666±0.003	0.713±0.001	0.547±0.006	0.624±0.001	0.514±0.008	0.329±0.002	0.279±0.002
Llama 3.1 (8b)	*max*	0.909	0.898	0.748	0.769	0.660	0.735	0.584	0.466	0.409
	min	0.376	0.300	0.435	0.629	0.166	0.456	0.320	0.252	0.175
	Original	0.809	0.759	0.660	0.749	0.494	0.676	0.480	0.426	0.237
	avg±var	0.832±0.009	0.805±0.011	0.668±0.003	0.737±0.001	0.496±0.013	0.652±0.003	0.488±0.003	0.387±0.002	0.282±0.002
Flan-T5-xl (3b)	*max*	0.940	0.960	0.809	0.776	0.743	0.706	0.761	0.685	0.496
	min	0.836	0.807	0.576	0.449	0.487	0.166	0.281	0.358	0.155
	Original	0.939	0.939	0.746	0.631	0.649	0.467	0.381	0.493	0.311
	avg±var	0.899±0.001	0.901±0.002	0.695±0.004	0.661±0.009	0.626±0.003	0.539±0.012	0.520±0.010	0.507±0.008	0.328±0.008

Table 4.2: *The maximum (max), minimum (min), average (avg), variance (var) of the model performance across different label names in the prompt for each validation set. The performance of the original label set (Original) is also included, showing that they fail to reach the maximum performance LLMs could get. The extent of the gap between the maximum and minimum performances is represented using colors:* $max-min > 0.3$, $0.2 < max-min < 0.3$. *The greater the variance, the greater the impact of label lexical choice, and the greater the potential utility of optimizing the label set.*

Lexical Choice As shown in Table 4.2, performance varies across datasets and models solely due to changes in the label names within the prompt. The variations are more pronounced than those reported in previous studies (Mu et al. 2024). The gap between the highest and lowest performance exceeds 0.1 for all datasets and models, and surpasses 0.2 on more than half of the datasets. We observe that certain stance labels could potentially trigger biased predictions, leading to extremely low performance. For example, Flan-T5 tends to always output *support* when using *support*, *deny*, *neither* for the semeval dataset, and Llama 3 overly predicts *con* when prompted by the label set *pro*, *con*, *neutral* for the emergent dataset.

[5]Potential impact of the decoding strategy can be found in the Appendix 4.7.4.

Label Order On average, we observe limited influence of label order (see the Appendix 4.7.5 for details). However, we are also interested in extremes of performance change caused by shifting label orders, particularly the maximum performance gain and drop, and whether the extreme gain or drop correlates with the lexicons used for the label names. Therefore, for each dataset, we first select the top-k[6] optimal and poor label sets based on their performance in the same order (i.e. the default order). We then examine the maximum increase or decrease of the performance after re-arranging the label orders for optimal and poor label sets, respectively.

We find that altering the order for the optimal label sets has the risk of high performance drops (e.g., even more than 0.2 on the binary classification snopes dataset, Figure 4.2a in Appendix). While performance improvements are possible, the gains are relatively limited (lower than 0.1 on all datasets). Conversely, re-arranging the order for the poorly performing label sets offers potential for substantial improvement (Figure 4.2b). This high improvements for certain label sets after re-ordering suggests that the poor performance may partly stem from the initial sub-optimal ordering.

Label Elaborations Similarly, we select the top-k[7] optimal and sub-optimal single-word label sets based on their performance, and examine the performance change after elaborations. We observe that the models are robust to the elaborations for either optimal or poor single-name label sets (full results in Table 4.9), indicating that adding task related details or increasing the label token lengths brings limited impact on average. However, we also observe relatively large performance change on certain datasets. Specifically, for the rumoureval dataset with Llama 3, we notice a performance drop larger than 0.2 when elaborating an optimal label set. Performance on the ibmcs dataset with Llama 3.1 could increase 0.2 when elaborating a poorly-performing label set.

4.3.4 Suggestions to Practitioners

Based on our results and analysis, we provide the following suggestions to practitioners in zero-shot ICL for classification:

1. Lexical designing for the label names should be considered as an important step in prompt engineering.

2. Single-word class label without elaboration on task information is able to achieve high performance in most cases, i.e. adding extra information does not yield better results.

3. If the practitioner has selected a set of optimal lexicons for label options based on a specific order, exploring alternative label orders can be redundant due to the limited performance gain brought from high computational costs. However, if a label set is chosen randomly, experimenting with different label orders may yield meaningful improvements (see Figure 4.2).

[6]Due to the exponential increasing of label order options and our limited computational resources, we set k=15 for binary classification; k=10 for three-way classification and k=2 for four-way classification

[7]k=15 for binary classification, k=30 for three/four-way classification.

4.4 Neuron Analysis for Label Selection

Although our findings indicate the importance of label word selection for text classification in zero-shot ICL, current studies lack consideration of this factor. Therefore, we conduct empirical analysis to gain insights into the underlying mechanism of lexical choice for single-word label names.

We preliminarily explored related approaches discussed in Section 4.2, including prompt perplexity (Gonen et al. 2023) and model internal representation of label words (Wang et al. 2023), but they did not yield any significant correlation with zero-shot ICL model performance (see the Appendix 4.7.9 and 4.7.10 for details). Meanwhile, various studies have indicated the correlation between neuron activation pattern in FFN and model performance (Kuzmin et al. 2023, Tang et al. 2024, Stolfo et al. 2024, Wu et al. 2024). Inspired by the finding that the presence of outliers in neural networks is predictive of quantization and pruning performance for the layers of LLMs (Kuzmin et al. 2023), we establish a new hypothesis: the model performance influenced by label names is correlated with the number of outliers in the neurons within FFN in the decoder of the LLMs. We empirically validate our hypothesis on the nine stance classification datasets, as well as two topic classification datasets (AG News (Zhang et al. 2015) and TweetTopic (Antypas et al. 2022)) to show the generalizability to other NLP tasks.

Methodology For each FFN module in layer i in the decoder, it can be denoted as follows:

$$h^i = (\text{act\_fn}(\tilde{h}^i W_1^i) \otimes \tilde{h}^i W_3^i) \cdot W_2^i. \tag{4.1}$$

where $\tilde{h}^i$ is the output hidden states from multi-head self-attention module. The activation function (act_fn) for Flan-T5 and Llama 3/3.1 is Gaussian Error Linear Unit (GELU) (Hendrycks & Gimpel 2016) and Sigmoid Linear Unit (SiLU) (Hendrycks & Gimpel 2016, Elfwing et al. 2018), respectively.

A *neuron* is defined as the linear transformation of each column in W_1^i followed by the activation function. Here, we study the last layer I's output of act_fn($\tilde{h}^I W_1^I$) (denoted as N_I) for the predicted first token of the label name in the model generation. Following Kuzmin et al. (2023), we measure the number of outliers over the neuron output distribution (N_I) through kurtosis, given by:

$$\text{Kurtosis}[N_I] = \frac{\mathbb{E}[(N_I - \mu)^4]}{(\mathbb{E}[(N_I - \mu)^2])^2} \tag{4.2}$$

where μ is the mean of N_I. For each dataset, we average the kurtosis scores over the validation set for each candidate label set. We then calculate the Spearman correlation between model performance and the averaged kurtosis score.

Results Table 4.3 shows that, for most datasets, there is statistically significant negative correlation between model performances and kurtosis scores across models and activation functions, indicating that fewer outliers in the neurons of the final layers are associated with enhanced zero-shot ICL performance. This observation implies that the kurtosis score of neuron activation distribution in FFN of the last decoder layer of LLMs could potentially serve as an effective signal for selecting optimal label names in zero-shot classification.

Model	Stance Classification									Topic Classification	
	perspectrum	ibmcs	snopes	scd	emergent	semeval	rumoureval	arc	vast	TweetTopic	AG News
Llama 3 (8b)	-0.4921*	-0.3787*	-0.4359*	-0.4217*	-0.5781	-0.2618*	-0.3764*	-0.1662	-0.3639*	-0.4994*	-0.2447*
Llama 3.1 (8b)	-0.4103*	-0.3642*	-0.1874	-0.0686	-0.4310*	-0.2232*	-0.3944*	-0.1708*	-0.1208	-0.2196*	0.0200
Flan-T5-xl (3b)	-0.4476*	-0.3638*	-0.6014*	0.2353	-0.1089	-0.2881*	-0.1714*	-0.5742	0.1003	0.1587	0.0398

Table 4.3: *Spearman correlation co-efficiency between model performance on validation set and kurtosis of neurons in the last layer. Mark with * when p value is lower than 0.05.*

4.5 LOADS: Label set Optimization via Activation Distribution kurtosiS

Motivated by the above observation that the fluctuated zero-shot ICL performance caused by different label names could be attributed to the number of outliers in neurons in the last layer of LLMs, we propose LOADS to obtain an optimal label set for a given classification task in a post-hoc setting.

4.5.1 Method

We design a three-step pipeline based on LOADS for automatic label selection in zero-shot ICL:

1. Create a list of candidate label sets for class options in the prompt (see Section 4.3). The label names in each set should follow the same order.
2. Rank the list of label options based on the kurtosis score of the neuron activation in FFN of the last decoder layer (averaged across the validation set).
3. Choose the label set with the lowest averaged kurtosis score.

The above LOADS-selected label set can then be used in the standard zero-shot ICL on test sets.

4.5.2 Evaluation

Setups We randomly sample 100 data points from the validation set for label selection and test the selected label sets on the official test sets.

- **Baselines:** We compare the model performance of using label words selected by LOADS to the following three approaches: (1) *Original label words*: we use the original label words from the dataset (i.e., the labels in Table 4.1), as it is the conventional and widely adopted practice; (2) *Original label words with a verbalizer*: after prompting the LLMs with the original label words, we employ our pool of candidate label words (See Section 4.3.1) as a verbalizer and incorporate their probabilities into the final probability of the same class; (3) *Self-generated label words*: we prompt the LLMs without providing any class options and select the candidate label words with the average highest probability at the first generated label token.
- **LLMs:** In addition to *Llama3* (8b), *Llama 3.1* (8b) and *Flan-T5-xl* (3b), we also examine whether LOADS could generalize to other model families and model sizes by including instruction-tuned *Gemma-2-it (9b)* and *Flan-T5-xxl* (13b).

Results Table 4.4 presents the zero-shot ICL model performances on stance classification and topic classification datasets with different label word selection strategies. The results demonstrate that employing LOADS to select label sets for zero-shot ICL prompts yields superior performance compared to other baseline approaches on most of the datasets. The improvement is consistent across NLP tasks and datasets, model architectures and sizes, as well as prompt templates[8].

Also, we observe limited benefits from adopting the verbalizer in post-processing, since LLMs tend to give the predicted label word tokens high probability in most of cases. Prompting with the self-generated label words rarely results in the best performance, while with the risk of leading to extremely low performance on certain datasets (e.g., perspectrum and snopes datasets with Llama 3).

Furthermore, potential data leakage could have significant impact on LOADS, indicated by the high performance achieved by the original label words on the AG News dataset with Flan-T5 models which was instruction-tuned with this dataset. It also aligns with the results in Table 4.3 where no statistically negative correlation was observed on the AG news dataset with Flan-T5-xl.

Model	Method	Stance Classification									Topic Classification	
		perspectrum	ibmcs	snopes	scd	emergent	semeval	rumoureval	arc	vast	TweetTopic	AG News
Llama 3 (8b)	LOADS	0.8431	0.7684	0.5516	0.6698	0.5870	0.6445	0.4097	0.3662	0.3190	0.7752	0.7660
	Original Label	0.8187	0.5485	0.4387	0.6504	0.3416	0.5565	0.3487	0.3130	0.2395	0.7742	0.7594
	Original + Verbalizer	0.8185	0.5485	0.4306	0.6578	0.3400	0.5576	0.3476	0.3131	0.2395	0.7742	0.7594
	Self-generated	0.6912	0.5802	0.3314	0.6607	0.3692	0.6032	0.2745	0.2837	0.3070	0.5851	0.6235
Llama 3.1 (8b)	LOADS	0.8789	0.8856	0.6212	0.7274	0.6403	0.6581	0.3642	0.3637	0.2458	0.7988	0.7306
	Original Label	0.8064	0.6783	0.4426	0.6983	0.4337	0.6492	0.3528	0.3784	0.2126	0.7909	0.7273
	Original + Verbalizer	0.8064	0.6783	0.4426	0.6983	0.4262	0.6448	0.3534	0.3838	0.2128	0.7909	0.7273
	Self-generated	0.6244	0.4918	0.5146	0.6798	0.2984	0.6421	0.3409	0.3373	0.2578	0.5956	0.6217
Gemma 2 (9b)	LOADS	0.9110	0.9247	0.6233	0.7595	0.5989	0.6784	0.4896	0.4858	0.3439	0.8032	0.8451
	Original Label	0.8966	0.8728	0.6454	0.7707	0.5704	0.6874	0.4778	0.4858	0.3383	0.8241	0.8316
	Original + Verbalizer	0.8966	0.8728	0.6424	0.7707	0.5523	0.6873	0.4689	0.4748	0.3383	0.8246	0.8322
	Self-generated	0.9017	0.9113	0.6656	0.7480	0.5833	0.6621	0.3549	0.4657	0.3178	0.6303	0.8215
Flan-T5-xl (3b)	LOADS	0.9334	0.9380	0.6881	0.5997	0.5813	0.4951	0.4759	0.4688	0.3857	0.8530	0.8556
	Original Label	0.9305	0.8971	0.7267	0.6341	0.5580	0.5697	0.3837	0.4628	0.3473	0.8071	0.9209
	Original + Verbalizer	0.9305	0.8953	0.7026	0.6507	0.5586	0.5712	0.3852	0.4613	0.3482	0.8091	0.9209
	Self-generated	0.7914	0.7384	0.3677	0.6974	0.3883	0.5954	0.2986	0.3343	0.3042	0.8501	0.7368
Flan-T5-xxl (13b)	LOADS	0.9428	0.9644	0.6905	0.6158	0.5938	0.5257	0.3852	0.6151	0.4218	0.7727	0.7742
	Original Label	0.9407	0.9630	0.7814	0.7598	0.5614	0.5697	0.2016	0.6065	0.3278	0.6643	0.9177
	Original + Verbalizer	0.9407	0.9621	0.7716	0.7598	0.5631	0.5705	0.2011	0.6062	0.3278	0.6643	0.9177
	Self-generated	0.8622	0.8384	0.4226	0.7315	0.4309	0.6182	0.3110	0.6115	0.2881	0.7242	0.9177

Table 4.4: *Comparison of zero-shot ICL performance on test sets between prompting with* LOADS*-selected label names versus the other three baseline approaches. We underline the highest model performance (statistically significant with paired chi-squared test).*

4.5.3 Analysis

Cross-lingual Transferability Previous research (Zhang, Li, Hauer, Shi & Kondrak 2023) has demonstrated that for non-English datasets, prompting with English task instructions (including label words) while keeping the input in the original language often yields superior performance than non-English instructions. Therefore, we investigate whether the optimal English label sets selected by LOADS on English dataset can also enhance performance for non-English datasets in this scenario.

We manually translate the English rumoureval Twitter test set into French and Portuguese. We select the optimal label set based on LOADS with 100 randomly sampled data from the

[8]The analysis of prompt sensitivity can be found in the Appendix 4.7.7.

English rumoureval validation set. We then prompt the LLMs with English task instructions and French/Portuguese inputs. Table 4.5 presents the results, indicating that the optimal English label set identified by LOADS also effectively improves performance on non-English datasets when the instruction is provided in English.

Model	Method	French	Portuguese
Llama 3 (8b)	LOADS	0.5020	0.4528
	Original Label	0.4544	0.4284
Llama 3.1 (8b)	LOADS	0.4728	0.4278
	Original Label	0.3731	0.3679
Flan-T5-xl (3b)	LOADS	0.5137	0.3912
	Original Label	0.4189	0.3317

Table 4.5: *Performance when LLMs are prompted with English instructions (including label options) and French/Portuguese inputs. Label options are selected by* LOADS *with English validation data.*

Data Efficiency To show the merit of data efficiency of LOADS, we randomly sample 50, 100, 300, 500 or 1000 data points from the validation set and compare the rankings of the label sets based on LOADS. Due to the resource restriction, we conduct experiments on snopes (binary classification) and emergent (three-way classification) datasets with Flan-T5-xl and Llama 3.

The results show that the rankings of the top 5 label sets remain consistent across different sample sizes. It suggests that LOADS can achieve comparable performance even with a smaller number of unlabeled data samples than 100, further highlighting the data efficiency of our proposed method. We provide computational cost estimation of LOADS using 100 unlabelled samples in Appendix 4.7.8.

Label Transferability We explore whether LOADS-selected label sets can be generalised across datasets or even models for NLP tasks such as stance classification where different label lexicons are used to represent the same classes across datasets. Specifically, for each dataset D_i on each LLM M_j, we select the optimal label set $L_{D_iM_j}$ through LOADS. To analyse whether the LOADS-selected label set on one dataset could be adapted to another related dataset with the same LLM M, we calculate the overlap of optimal labels ($L_{D_i\_M}$) between each dataset. Similarly, we examine the overlap of optimal labels ($L_{D\_M_i}$) between each model to explore whether the LOADS-selected optimal labels for dataset D could be adapted across LLMs. We only focus on the positive and negative stances to enable comparison across binary, three-way and four-way stance classification datasets.

Our results indicate that LOADS-selected label sets is transferable across datasets on the same LLM, highlighting the potential of leveraging LOADS to identify optimal label words with established related datasets, avoiding the need to collect samples for the target new dataset. For example, the positive-negative stance label pairs identified for Llama 3 is *endorse* and *deny* across all the stance classification datasets. However, we find that the label words selected for a specific dataset on one LLM often differ from those identified for another LLM, suggesting LOADS' dependency on the underlying model architectures and parameters.

In summary, the LOADS-selected label sets tend to be model-dependent rather than

dataset-dependent. This observation aligns with the mechanism of LOADS, as the neurons and their distributions are inherently tied to the specific model. We hypothesize that this may suggest a correlation between the LOADS-selected label words and the LLMs' internal representation or understanding of the target NLP task or concept (e.g., what is stance), highlighting potential directions for future studies.

4.6 Conclusion

We study the impact of label options in the prompt for classification in zero-shot ICL, including lexical choice, label order, and label elaborations. We observe a significant effect of the lexicons used to represent label words in the prompt, also linking to the models' sensitivity to the label order. Through neuron activation analysis, we find that optimal label sets produce fewer outlier neurons in LLMs' feed-forward networks. We then propose LOADS, a novel method for selecting optimal label sets using activation distribution kurtosis. Prompting with LOADS-selected label sets consistently outperforms the use of original dataset labels across different models. Our approach is post-hoc, data-efficient and requires no gradient propagation or model fine-tuning. It also demonstrates cross-lingual transferability when using English instructions for non-English datasets. By showing that carefully selecting label sets based on neuron activation patterns can significantly enhance model performance without requiring additional training or labeled data, this paper has important implications for leveraging LLMs in zero-shot classification.

Limitations

Our experiments focused primarily on stance classification tasks. We chose this task because the label ambiguity is an identified challenge (i.e., label names could be replaced by a sufficient number of synonyms without altering their meanings and scopes in the original study), and it has sufficient datasets for empirical study. Although we have tested the generalisabity of our findings on topic classification, with more datasets released and new tasks proposed in future studies, studies could be conducted to explore whether our findings generalize to a broader range of classification tasks and domains. Also, although we examined multiple models from the Flan-T5 and Llama families, our study did not include other popular language models (such as Phi[9] or Mistral [10]) due to computational resource limitation. Expanding the range of models would provide a more comprehensive understanding of label option's impact across different architectures.

Another limitation of our study is English-language bias. Although we have explored the cross-lingual transferability to French and Portuguese on one dataset, more extensive multilingual testing is needed to ensure the approach's effectiveness across diverse languages and cultures.

Our method, while more efficient than gradient-based approaches, still requires running inference on a subset of data to compute activation statistics. This may be challenging for resource-constrained environments or very large models. The efficiency of our method may also be challenged when the classification task contains a very large number of class categories.

[9] `https://huggingface.co/microsoft/phi-2`

[10] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`

Furthermore, we ensure the inclusion of samples for each class. The effectiveness of our method might vary when the validation set is highly imbalanced or even lack of data for the minority class. The effectiveness of different distribution metrics is out of scope but we acknowledge that it may have significant improvement for our method.

Lastly, while we focused on technical performance, future work should consider potential biases introduced by label set choices and their implications for fairness and inclusivity in classification tasks. Addressing these limitations in future research will help to further validate and refine our approach to optimal label set selection for zero-shot ICL.

Acknowledgments

This work was partially funded by EMIF managed by the Calouste Gulbenkian Foundation[11] under the "Supporting Research into Media, Disinformation and Information Literacy Across Europe" call (ExU – project number: 291191)[12] and the UK's innovation agency (InnovateUK) grant number 10039039 (approved under the Horizon Europe Programme as VIGILANT[13], EU grant agreement number 101073921). Yue Li was supported by a Sheffield–China Scholarships Council PhD Scholarship. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.

4.7 Appendix

4.7.1 Datasets

We summarise the datasets we used in our study in Table 4.6. For the stance classification datasets without official validation sets, we use the train/validation splits provided by Schiller et al. (2021). We utilize the official test set for AG news[14], and the official "train/test random split in the COLING 2022 paper" for TweetTopic dataset[15]. TweetTopic dataset has six class categories, potentially resulting in more than 4,000 different label sets if we consider only five synonymy words for each category (i.e., more than 12,000 experiments on three LLMs). Due to our limited computational resource, we experiment with three topics: *pop culture*, *daily life*, and *science & technology*.

4.7.2 Data Leakage

As far as we know, Llama 3 and Llama 3.1 are not supervised fine-tuned with any public stance classification datasets. Flan-T5 is fine-tuned on Super-NaturalInstructions dataset (Wang et al. 2022), containing two English stance classification tasks (Kobbe et al. 2020) (i.e., task 209 and 513). The tasks are formed as a binary ("in favor" and "against") and a three-

[11] The sole responsibility for any content supported by the European Media and Information Fund lies with the author(s) and it may not necessarily reflect the positions of the EMIF and the Fund Partners, the Calouste Gulbenkian Foundation and the European University Institute.

[12] `exuproject.sites.sheffield.ac.uk`

[13] `https://www.vigilantproject.eu`

[14] `https://huggingface.co/datasets/sh0416/ag_news`

[15] `https://huggingface.co/datasets/cardiffnlp/tweet_topic_single`

Dataset Name	Source	# of Label Sets
scd	Debates	31
perspectrum	Debates	31
snopes	News	31
ibmcs	Debates + Wikipedia	31
vast	Debates + Artificial	62
emergent	News	62
semeval	Social Media	93
rumoureval	Social Media	248
arc	Debates	62
AG News	News	50
TweetTopic	Social Media	64

Table 4.6: *Datasets and the number of label sets we experiment with for each dataset.*

way ("in favor", "against" and "neutral") classification, respectively. There is no overlapping between these two datasets and our nine experimented datasets.

4.7.3 Label Pool Creation

Stance Classification Following the pipeline we described in Section 3 of the main paper, we collect the seed label sets from the nine stance classification datasets (see Table 1). For the semeval dataset, the label set in the original paper ("favor, against, neither" in Table 1 in main paper) is slightly different from the set used in their published dataset ("favor, against, none"), so we consider both of them.

For positive and negative stance label names, we aim to acquire word-pairs with semantically opposite meaning. We first extract antonym for each positive and negative seed stance label from WordNet. Since we obtain limited antonyms in this way, Claude is then used to generate synonym for each seed positive-negative stance label pairs. An example of the prompt we used is: *Provide 5 different pairs of synonyms for "support" and "deny". They are supposed to be labels for stance classification.* We use WordNet to obtain synonyms for the rest of stance labels if there are any. For the label names that represent "neutral" stance in the original study, such as "observing" and "comment", we take "neutral" as their synonyms. Finally, we manually select the appropriate label names generated by Claude. The number of label sets we experiment with for each dataset is listed in Table 4.6.

Topic Classification Similarly, we follow the pipeline to collect and generate synonyms for each topic category. For TweetTopic dataset, since *pop culture* is a mixture of multiple sub-topics as discussed by Antypas et al. (2022), we also consider the synonyms of the sub-topics. We use every possible combination of synonyms among topic categories for TweetTopic dataset. For AG news, there are total 160 combinations. We randomly sample 50 of them due to limited computational resources. The number of label sets we experiment with for two datasets is listed in Table 4.6.

4.7.4 Decoding Strategies

We adjust the temperature (0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4) used for sampling-based decoding and compare their performances with the greedy search based performance for emergent and snopes dataset on Flan-T5-xl and Llama 3.

A temperature value larger than 1.0 – flattening the probability distribution – tends to harm the performance especially for Flan-T5-xl, which generates outputs irrelevant to stance. When temperature is lower than 1.0, introducing randomness in decoding through sampling may benefit the performance, but not significantly (in most of cases improvement is lower than 0.07). We summarise the maximum performance increase or decrease comparing with greedy search in Table 4.7.

Model Name	Temperature	snopes		emergent	
		+	-	+	-
Flan-T5	0.2	0.021	0.049	0.037	0.049
	0.4	0.055	0.066	0.066	0.068
	0.6	0.033	0.099	0.034	0.064
	0.8	0.033	0.145	0.066	0.107
	1.0	0.022	0.189	0.036	0.149
	1.2	0.037	0.264	0.046	0.196
	1.4	0.015	0.428	0.042	0.361
Llama 3	0.2	0.050	0.063	0.068	0.035
	0.4	0.044	0.047	0.073	0.073
	0.6	0.040	0.062	0.075	0.059
	0.8	0.027	0.062	0.066	0.107
	1.0	0.054	0.086	0.100	0.103
	1.2	0.038	0.072	0.069	0.171
	1.4	0.030	0.104	0.128	0.142

Table 4.7: *The maximum performance increase (+) and decrease (-) if adopting sampling-based decoding rather than greedy search.*

Prompting with Label Explanation

We investigate whether the performance variance caused by different lexical choices of the label names could be mitigated or lowered by including the explanation of the label names in the prompt. We experiment with emergent and snopes datasets on Flan-T5-xl and Llama 3. We add the following class explanations in the prompt template after the class options for snopes and emergent datasets respectively: (1) snopes: *If the text supports that claim, answer with "{positive stance}"; if the text opposes the claim, answer with "{negative stance}"*; (2) emergent: *If the headline supports the claim, answer with "{positive stance}"; if the headline opposes the claim, answer with "{negative stance}"; if the claim is discussed in the headline but without assessment of its veracity, "{neutral stance}"*.

We observe that including these label name explanations in the prompt may help with the label sets that achieve the lowest zero-shot performance. As for the snopes dataset, its worst performance would increase from 0.5400 to 0.555 (Llama 3, labels: *supportive* and *opposed*) or from 0.5766 to 0.6568 (Flan-T5-xl, labels: *for*, *against*). As for emergent, its lowest performance would increase significantly from 0.3877 to 0.5600 with Llama 3 (labels: *pro*, *con* and *neutral*). However, when using Flan-T5-xl, the inclusion of the class explanation even decrease the worst performance from 0.4870 to 0.3775 (labels: *support*, *deny* and *neutral*).

More importantly, the benefits from label explanations in the prompt would not close the gap between the optimal and sub-optimal label sets, comparing the above improved performance with the maximum performances in Table 2 in the main paper.

4.7.5 Label Order Results

We present the averaged absolute performance difference after re-ordering the label names in the prompt in Table 4.8. The influence is limited on average.

Dataset	Flan-T5	Llama 3	Llama 3.1
perspectrum	0.0148	0.0372	0.0771
ibmcs	0.0195	0.0696	0.0961
snopes	0.0296	0.0772	0.1259
scd	0.0309	0.0288	0.0484
emergent	0.0211	0.0689	0.1578
semeval	0.0230	0.0304	0.0527
vast	0.0311	0.0465	0.0495
rumoureval	0.0355	0.0720	0.0439
arc	0.0152	0.0606	0.0612

Table 4.8: *The average absolute performance change after re-ordering the label options in the prompt.*

The maximum performance gain and drop on each dataset after re-ordering the label names for the top-k optimal and poor label sets with Llama3, Llama 3.1 and Flan-T5-xl are in Figure 4.2.

4.7.6 Label Elaboration Results

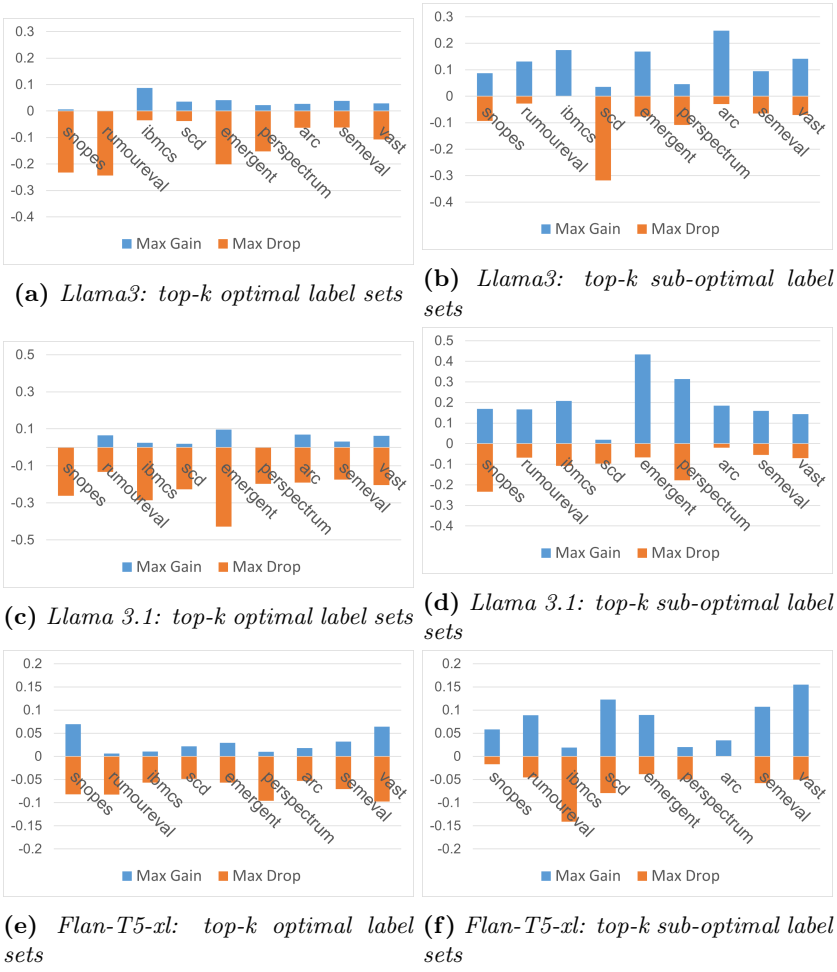
We supplement the averaged absolute performance difference for each level of elaboration on Llama 3.1 and Flan-T5-xl in Table 4.9.

	Dataset	E_1		E_2		E_3	
		Opt.	Sub-opt.	Opt.	Sub-opt.	Opt.	Sub-opt.
Llama 3	perspectrum	0.016	0.018	0.010	0.015	0.017	0.009
	ibmcs	0.027	0.041	0.024	0.014	0.029	0.044
	snopes	0.055	0.040	0.054	0.026	0.029	0.018
	emergent	0.051	0.053	0.047	0.038	0.022	0.040
	rumoureval	0.095	0.036	0.084	0.020	0.058	0.102
	arc	0.017	0.024	0.015	0.021	0.035	0.049
Llama 3.1	perspectrum	0.027	0.020	0.048	0.025	0.037	0.022
	ibmcs	0.033	0.018	0.054	0.031	0.057	0.067
	snopes	0.015	0.021	0.032	0.020	0.034	0.033
	emergent	0.048	0.077	0.048	0.107	0.034	0.133
	rumoureval	0.041	0.040	0.037	0.032	0.039	0.035
	arc	0.032	0.046	0.029	0.036	0.042	0.046
Flan-T5-xl	perspectrum	0.009	0.026	0.010	0.021	0.013	0.021
	ibmcs	0.014	0.015	0.016	0.020	0.017	0.028
	snopes	0.020	0.026	0.022	0.032	0.033	0.016
	emergent	0.026	0.025	0.044	0.031	0.042	0.042
	rumoureval	0.059	0.040	0.060	0.031	0.086	0.018
	arc	0.034	0.042	0.025	0.038	0.031	0.041

Table 4.9: *The average absolute performance change after elaborating for optimal or poor single-word label sets with Llama 3.1 and Flan-t5-xl (E_1, E_2, E_3 see Figure 1 in main paper).*

4.7.7 Prompt Sensitivity Analysis of LOADS

To analyse the the prompt sensitivity of LOADS, we test it on different prompt templates, select the label set through LOADS, and then compare the performance with that on the label sets in the original dataset.



(a) *Llama3: top-k optimal label sets*

(b) *Llama3: top-k sub-optimal label sets*

(c) *Llama 3.1: top-k optimal label sets*

(d) *Llama 3.1: top-k sub-optimal label sets*

(e) *Flan-T5-xl: top-k optimal label sets*

(f) *Flan-T5-xl: top-k sub-optimal label sets*

Figure 4.2: *The maximum performance gain (positive value) and drop (negative value) on each dataset after re-ordering the label names for the top-k optimal and sub-optimal label sets with Llama3, Llama 3.1 and Flan-T5-xl.*

Due to the computational resource constraints, we manually craft two prompts and test LOADS with the four binary stance classification datasets on Llama 3. In the two prompts, we replace *Given a [text1_name] and a [text2_name], detect the stance that the [text2_name] has towards the [text1_name]* (see Section 3.2 in main paper) with two different queries:

1. Prompt 1: *What is the stance of [text2_name] towards [text1_name]?*
2. Prompt 2: *What stance does [text2_name] take regarding [text1_name]?*

As shown in Table 4.10, although different prompts with the same label sets may result in performance changes as expected (compare with Table 5 in main paper), LOADS is robust to different prompts used for the label selection. The performance gap between LOADS-selected and original label sets tends to be similar across prompt templates.

	Dataset	LOADS	Original Label
Prompt 1	snopes	<u>0.5926</u>	0.4984
	ibmcs	<u>0.8737</u>	0.7303
	perspectrum	<u>0.8925</u>	0.8658
	scd	<u>0.6895</u>	0.6860
Prompt 2	snopes	<u>0.6191</u>	0.5336
	ibmcs	<u>0.8619</u>	0.7523
	perspectrum	<u>0.8921</u>	0.8619
	scd	0.6836	<u>0.6931</u>

Table 4.10: *Performance comparison on Llama 3 when using* LOADS*-selected label sets (lowest kurtosis) and using original label sets (original label) with prompt 1 or prompt 2. The higher performance is underlined.*

4.7.8 Computational Cost Estimation

Following previous work (Kaplan et al. 2020, Liu, Tam, Mohammed, Mohta, Huang, Bansal & Raffel 2022), we estimate that a decoder-only LLM with N parameters uses $2N$ FLOPs per token for inference. We suppose that: (1) the input token length for the dataset we are interested in is L on average; (2) the number of candidate label sets is X; (3) 100 unlabelled texts are used for LOADS. Therefore, the total FLOPs taken by LOADS would be $2N * L * X * 100 = 200NLX$.

4.7.9 Perplexity Analysis

As discussed in Section 2 in main paper, Gonen et al. (2023) empirically show that zero-shot ICL performance is statistically negative correlated with the perplexity of the prompt with input. However, they did not take into account the label options in the prompt when calculating the perplexity. Therefore, we further investigate whether the perplexity is also correlated with the variance zero-shot ICL performance caused by different label names.

Specifically, we use the prompt template in Section 3.2 in main paper, and calculate the perplexity of prompts with inputs and different label sets. Following Gonen et al. (2023), for each label set, we average the perplexity over the dataset. And then we adopt spearman correlation test between the averaged perplexity scores and model performances. Due to the computational restriction, we experiment with all the binary datasets on Flan-T5-xl and Llama3-8b. Since Flan-T5 is an encoder-decoder model where perplexity has a loose definition, we treat the encoder input as an empty string when calculating perplexity.

The results in Table 4.11 indicate that there is no statistically significant correlation between prompt perplexity and model performance if considering different label sets in the prompt.

		perspectrum	ibmcs	snopes	scd
Llama 3	*coefficient*	0.0068	0.1641	0.0394	0.1698
	p value	0.9707	0.3774	0.8303	0.3608
Flan-T5	*coefficient*	0.0738	0.1733	-0.0500	-0.2273
	p value	0.6929	0.3511	0.7892	0.2187

Table 4.11: *Spearman correlation between model performance and prompt perplexity. P-values are all larger than 0.05, indicating no statistical significance.*

4.7.10 Label Attention Key Similarity Analysis

In this section, we explore whether the closely related observation on few-shot ICL could be directly adopted to zero-shot ICL. Specifically, we focus on the study discussed in Section 2 in main paper, where Wang et al. (2023) suggest that when the LLM is prompted by demonstration with examples in a few-shot ICL setting, the model is likely to confuse the label categories if their key vectors in the attention modules are similar to each other.

Since this finding is easier to be tested on binary datasets, we experiment with the binary datasets on Llama 3 and Flan-T5-xl. We extract the key vectors in the attention module in each layer for each label name in the prompt. Then we calculate the cosine similarity between the vectors of two label names. Finally, we use spearman correlation test between similarity scores and model performances. As shown in Table 4.12, we do not observe statistically significant correlation between model performance and label key vector similarities.

		perspectrum	**ibmcs**	**snopes**	**scd**
Llama 3	*coefficient*	-0.0181	-0.0051	-0.2791	-0.1696
	p value	0.9244	0.9784	0.1283	0.3702
Flan-T5-xl	*coefficient*	-0.0595	0.0223	-0.0209	-0.0992
	p value	0.7503	0.9047	0.9108	0.5953

Table 4.12: *Spearman correlation between model performance and label's key vector similarity.*

4.7.11 Layer-Wise Output Projections Analysis

We hypothesize that LLM may jump to the output prediction at last layers when a sub-optimal label set is used in the prompt. Therefore, we extract the hidden states from each decoder layer and project them on the model vocabulary, so that we obtain the ranked position of the final predicted label's token in each layer (Elhage et al. 2021, Geva et al. 2022).

We observe that the hypothesis indeed holds in certain cases. We show an example on rumoureval dataset where we compare the averaged rank of the correctly predicted label *comment/neutral* in each decoder layer when Flan-T5-xl is prompt to choose from *support, deny, query, comment* or *support, deny, query, neutral*. Label set *support, deny, query, comment* performs worse than the set *endorse, deny, query, neutral* on this dataset. As shown in Figure 4.3, when using the relatively optimal label set *endorse, deny, query, neutral*, the rank of the final predicted label tends to move closer to the top at an earlier stage.

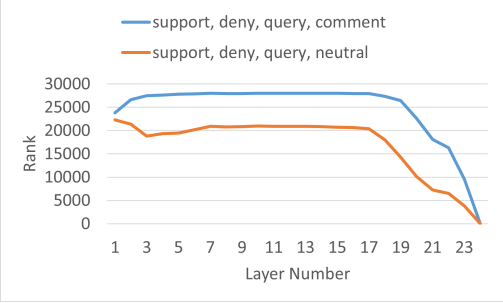


Figure 4.3: *The rank of the final correctly predicted label (comment orneutral) when Flan-t5-xl is prompted with two different label sets for rumoureval dataset.*

4.7.12 Human Translation Details

To translate the English Twitter rumoureval test set into French and Portuguese, we recruit volunteer students from translation studies in Brazilian and French universities. The students are given gift vouchers (0.6 pounds per tweet). Consent has been obtained from the students and our study has received approval from the Ethics Committee of our university.

We instruct the students to translate the tweets accurately, and preserve the original meaning, context, and tone of the tweet. They are also encouraged to leave notes for their translations. The translations are finished on Google Sheets.

Chapter 5

Publication IV: SCRum-9: Multilingual Stance Classification over Rumours on Social Media

Overview Previous chapters primarily investigate generalisation and adaptation of rumours stance classification for English. However, rumours circulate across social media platforms in a wide range of languages. Limiting research to English constrains the applicability of existing models, and risks overlooking linguistic and cultural variations that may substantially affect stance expression. Therefore, in the publication presented in this chapter, we introduce the current largest multilingual rumour stance classification dataset and systematically benchmark recent advanced approaches for language adaptation with LLMs, including ICL with language alignment and supervised training with LLM-generated multilingual data.

SCRum-9: Multilingual Stance Classification over Rumours on Social Media

Yue Li, Jake Vasilakes, Zhixue Zhao, Carolina Scarton

School of Computer Science, University of Sheffield, UK

Abstract

We introduce **SCRum-9**, the largest multilingual **S**tance **C**lassification dataset for **Rum**our analysis in **9** languages, containing 7,516 tweets from X. SCRum-9 goes beyond existing stance classification datasets by covering more languages, linking examples to more fact-checked claims (2.1k), and including confidence-related annotations from multiple annotators to account for intra- and inter-annotator variability (Figure 5.1). Annotations were made by at least two native speakers per language, totalling more than 405 hours of annotation and 8,150 dollars in compensation. Further, SCRum-9 is used to benchmark five large language models (LLMs) and two multilingual masked language models (MLMs) in In-Context Learning (ICL) and fine-tuning setups. This paper also innovates by exploring the use of multilingual synthetic data for rumour stance classification, showing that even LLMs with weak ICL performance can produce valuable synthetic data for fine-tuning small MLMs, enabling them to achieve higher performance than zero-shot ICL in LLMs. Finally, we examine the relationship between model predictions and

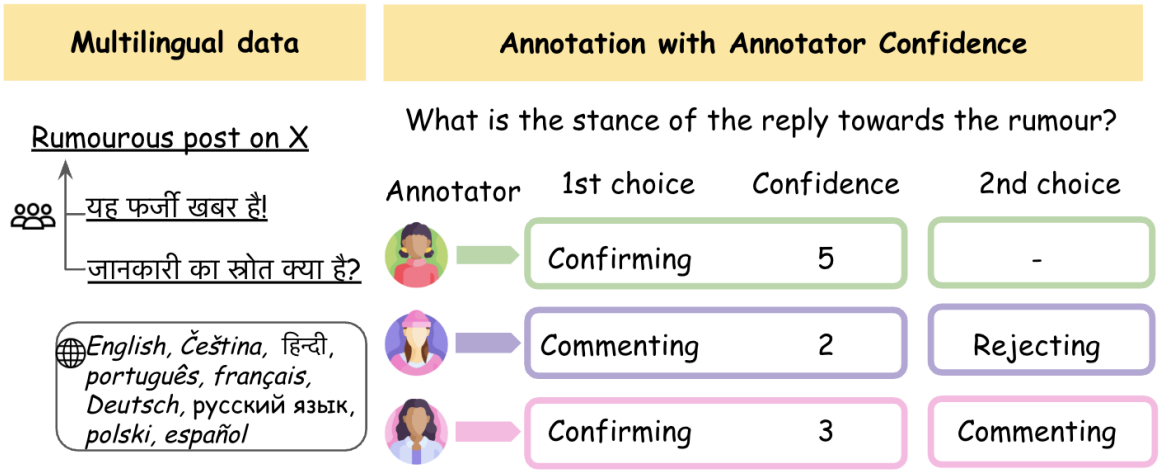


Figure 5.1: *An illustration of the multilinguality and annotation design of SCRum-9.*

human uncertainty on ambiguous cases finding that model predictions often match the second-choice labels assigned by annotators, rather than diverging entirely from human judgments. SCRum-9 is publicly released to the research community with potential to foster further research on multilingual analysis of misleading narratives on social media.

5.1 Introduction

Social media amplifies the spread of information, including dis- or misinformation and rumours, which often cross linguistic and geographic boundaries Shu et al. (2020). Rumour stance classification Zubiaga et al. (2018*c*) plays a key role in rumour analysis on social media by identifying whether users support, refute, or question a rumour. However, most existing datasets (Derczynski et al. 2017, Gorrell et al. 2019*a*) focus exclusively on English, despite the multilingual nature of online rumours. Efforts have been made to construct multilingual datasets for generic stance classification (Zotova et al. 2020, Vamvas & Sennrich 2020, España-Bonet 2023, Agerri et al. 2021, Hamdi et al. 2021), whose task framing, label granularity, and evaluation protocol are fundamentally different from stance classification for rumour or misinformation analysis (Hardalov et al. 2022*b*, Scarton et al. 2020*a*).

This paper introduces SCRum-9, the first large-scale multilingual benchmark dataset for rumour stance classification. The dataset consists of 7,516 reply tweets across nine diverse languages: Czech (CS), German (DE), English (EN), Spanish (ES), French (FR), Hindi (HI), Polish (PL), Portuguese (PT), and Russian (RU). SCRum-9 covers a broad range of topics with 2,156 distinct rumours collected from fact-checking websites. Each reply is manually annotated using established rumour stance classification schemes (Gorrell et al. 2019*a*), with an additional design (Figure 5.1) allowing annotators to provide a second-choice label when unsure, which aims to model annotator uncertainty (Mu et al. 2023). To the best of our knowledge, SCRum-9 is the most topically and linguistically diverse rumour stance classifica-

Dataset	Size	#Stance	#Lang	#Fact-Checks	Confidence Annotation	Released Annotations
PHEME (Zubiaga et al. 2016)	5.8k	4	2	5	✗	Aggregated
Stanceosaurus (Zheng et al. 2022)	28k	5	3	250	✗	Aggregated
Stanceosaurus 2.0 (Lavrouk et al. 2024)	32k	5	2	291	✗	Aggregated
SCRum-9 (Ours)	**7.5k**	**4**	**9**	**2,156**	✓	**Raw & Aggregated**

Table 5.1: *Comparison between SCRum-9 and existing multilingual stance classification datasets for misinformation or rumour analysis.*

tion dataset to date[1]. We also release all raw annotations to facilitate studies on annotators' uncertainty and subjectivity in multilingual stance classification.

We benchmark five open-source multilingual LLMs using ICL and observe substantial performance disparities across English and relatively low-resource languages included in SCRum-9. We conduct extensive analyses of strategies for improving ICL performance in non-English, including (1) direct machine translating non-English target input into English, or English demonstration examples into target non-English; and (2) incorporating language alignment signals into the prompt. We also evaluate two MLMs fine-tuned on English training data, machine-translated multilingual data, or synthetic multilingual data generated by LLMs.

Our key findings are as follows:

- Machine translating target input or demonstration examples is an effective and competitive strategy to improve ICL performance on relatively high-resource languages whose translation quality is reliable.
- Zero-shot ICL significantly outperforms or matches the performance of MLMs fine-tuned with English or machine-translated multilingual data.
- MLMs fine-tuned with LLM-generated multilingual data demonstrate promising results which are comparable to or even better than zero-shot ICL performance with the same LLM, while requiring substantially lower computational costs.
- In ambiguous cases where annotators provide second-choice labels, model predictions sometimes align with these alternatives, suggesting that the outputs mirror human uncertainty rather than being purely errors.

5.2 Related Work

5.2.1 Multilingual Stance Classification Datasets

Efforts have been made to construct multilingual stance classification datasets, such as CIC (Zotova et al. 2020), Xstance (Vamvas & Sennrich 2020), PoliOscar (España-Bonet 2023), VaxxStance (Agerri et al. 2021) and NEWSEYE (Hamdi et al. 2021). However, these datasets are typically centred around political opinions or social issues, rather than for misinformation or rumour analysis. Stance classification for the latter purpose, such as rumour stance classification, is often framed with distinct and more nuanced stance categories to support

[1]To be released under a CC BY-NC-SA 4.0 license upon acceptance.

fact-checking or rumour verification (such as the *questioning* stance to identify whether a related information is asked for verification) (Zubiaga et al. 2018*c*, Hardalov et al. 2022*c*). The differences in task framing, label granularity, and annotation guidelines render those general stance classification datasets not applicable to model development and evaluation in disinformation or rumour analysis scenarios.

Language Coverage There are only three multilingual stance classification datasets established for disinformation or rumour analysis as presented in Table 5.1, including PHEME (English and German), Stanceosaurus (English, Hindi, and Arabic) and Stanceosaurus 2.0 (Russian and Spanish), covering five non-English languages in total. Our proposed SCRum-9 dataset expands the linguistic coverage by incorporating four additional non-English languages not present in the above previous work: Polish, Czech, French, and Portuguese.

Annotation Protocol Stance classification is inherently subjective, as annotators may interpret the same post differently due to linguistic ambiguity, different cultural background, or personal judgments (Mu et al. 2023, Wu et al. 2023*b*). Existing English (Derczynski et al. 2017, Gorrell et al. 2019*b*) and multilingual (Zubiaga et al. 2016) rumour stance classification datasets release the majority-voted labels among multiple annotators as the final gold standard labels. However, the RumourEval 2017 dataset contains highly similar tweets with different aggregated stance labels (Derczynski et al. 2017, Li et al. 2019, García Lozano et al. 2017), illustrating inconsistencies and limitation that arise from relying solely on majority voting. Different from prior work, our annotation protocol explicitly encourages annotators to indicate their confidence level and provide a secondary label when uncertain. We release all raw annotation data from each annotator to enable future research on annotation uncertainty and subjectivity.

5.2.2 Multilingual Rumour Stance Classification

Several studies have explored multilingual or cross-lingual stance classification (Hardalov et al. 2022*a*, Zhang, Yang & Mao 2023, Scarton & Li 2021, Zheng et al. 2022, Barriere et al. 2022), but they mainly focus on general stance classification, which differs significantly in task framing and evaluation protocol from rumour stance classification (Scarton et al. 2020*b*). Furthermore, prior work has experimented with multilingual MLMs (e.g., multilingual BERT (Scarton & Li 2021)), with the focus on transferring knowledge learnt from the English source stance classification dataset to the target-language dataset. However, current LLMs (Grattafiori et al. 2024, Le Scao et al. 2023) have offered support for many relatively high-resource languages (e.g., the eight non-English languages in SCRum-9). The multilingual capability of LLMs on stance classification, especially rumour stance classification, and their potential to generate non-English synthetic data for those languages in multilingual learning remain under-explored.

5.3 The SCRum-9 Multilingual Dataset

We construct our dataset from rumours collected on fact-checking websites. To ensure both topic consistency and diversity across languages, we apply a topic-based filtering strategy when

selecting rumours for annotation. Our annotation scheme follows PHEME and RumourEval datasets (Derczynski et al. 2017, Gorrell et al. 2019*a*), while introducing additional design to explicitly capture annotator uncertainty inspired by Mu et al. (2023). Notably, SCRum-9 is different from existing rumour stance classification datasets that either focus on emerging rumours not known a priori (e.g., PHEME and RumourEval 2017 (Derczynski et al. 2017)) or restrict to a single topic (e.g., natural disasters in RumourEval 2019 (Gorrell et al. 2019*a*)).

5.3.1 Data Collection

We collect fact-checked claims and corresponding X posts from two sources:

- **The Database of Known Fakes (DBKF):** DBKF is a publicly available database of existing fact-checks and associated metadata from trusted fact-checking organisations.[2] It contains fact-checked claims, links to the fact-checking articles, as well as links to news and social media posts related to the claim, including those from X. The fact-checks included in DBKF cover a wide range of languages, including those listed above. We query DBKF for all data points in one of the nine target languages that are linked to one or more X posts.
- **Fact-Checking Websites:** We identify different trusted fact-checking websites for each language (see Table 5.3 in the Appendix for specific websites). With permission, we develop web scraping tools that crawl each site, identify all fact-check articles, and extract the claim and all associated X links.

The result of this step is a collection of claims linked to X tweet URLs. We use the X API[3] to obtain each tweet's text, replies, and other metadata. Given the large number of replies that many posts have, obtaining all replies would quickly reach the 1 million post limit enforced by the X API. We opt to only collect the tweets directly replying to the rumourous source tweet, as the responses in extended conversations become less informative and/or gradually shift towards topics less relevant to the rumour (Kochkina & Liakata 2020). We sample up to 60 direct replies for each post, keeping posts that attract sufficient public engagement (i.e., with more than 25 replies) and informative replies that do not contain only URLs, user mentions, or emojis.

5.3.2 Topic-Based Tweet Filtering and Pre-Processing

To ensure topic diversity but also mitigate models leveraging event-specific biases in model evaluation, we employed the multilingual BERTopic model (Grootendorst 2022), alongside manual review, to identify a set of topics that occur across languages, resulting in 63 coherent topics (details can be found in the Appendix). The topics range from broad ones such as COVID-19 and natural disasters, to specific rumours, e.g., that Switzerland has outlawed mammograms. We keep only rumourous source tweets that are assigned to these 63 topics by the topic model, except for Hindi, Czech, and Russian, where filtering using the topic model would severely reduce the total number of examples. After filtering, 53 of the original 63 topics were represented. Finally, we sample up to 1,500 tweet-reply pairs per language for

[2] `https://dbkf.ontotext.com`

[3] `https://developer.x.com/en/docs/x-api`

annotation such that we maximise the total number of topics represented. Statistics of the source tweets, replies, and fact-checked claims that were collected, filtered, and annotated are presented in Table 5.4 in the Appendix.

Before annotation, the tweet-reply pairs were anonymised by replacing all user `@` mentions with the string `@USER` and all URLs within tweets were replaced with `HTTPURL`. We discuss ethical considerations related to data anonymity in the Appendix.

5.3.3 Data Annotation

The annotation task is to determine the stance of a reply tweet towards its rumourous source tweet within a conversation thread on X. Stance is defined as the way in which the reply tweet's author regards the source tweet. Following PHEME and RumourEval datasets (Gorrell et al. 2019*a*), there are four possible stances [4]:

- **Confirming** - The author expresses agreement with source tweet.
- **Rejecting** - The author expresses disagreement with or denial of the source tweet.
- **Questioning** - The author asks for additional evidence or confirmation of the source tweet.
- **Commenting** - The author comments on the source tweet but does not take a clear stance. This includes cases in which the reply is unrelated to the source tweet.

Since SCRum-9 focuses on text-based rumour stance classification, annotators are also instructed to label replies that rely exclusively on the attached images or videos to interpret the stance as *Only refers to image/video*. These instances are excluded from the final dataset.

Confidence-Based Annotation Annotators are instructed to provide one or two stance annotations for each tweet-reply pair, based on their annotation confidence. Specifically, each annotator indicates a first-choice stance label, accompanied by a confidence rating on a 5-point Likert scale (1 = extremely uncertain, 5 = absolutely certain). If the confidence rating is lower than three, annotators are required to provide a second-choice stance label. To account for cases where no specific second label can be reasonably assigned, we include *Highly Uncertain* as an additional option. Detailed annotation guidelines are provided in the Appendix.

Annotator Recruitment and Quality Control

Annotators are native speakers of each target language from universities and research institutions in our network. The native English speakers were predominantly from the UK, and the native Portuguese speakers were from Brazil. To comply with our ethical policy, annotators received an information sheet, outlining the details regarding their participation and needed to sign a consent form.

[4] In this chapter, "confirming", "rejecting", "questioning" and "commenting" are used to refer to the same underlying taxonomy (i.e., "support", "deny", "query" and "comment") as in previous chapters. These terminologies are modified according to suggestions by our annotators.

Participants received a 40-minute online training session prior to beginning the annotation task, where we introduced the annotation guidelines, process, and interface. After the training session, participants completed quality control annotations on 60 held-out examples in English, which had been annotated by three researchers experts in the task. We selected participants who reached a high level of agreement with the researchers' annotations.[5] Annotators were compensated at the rate of $20 (USD) per hour (or equivalent in their currency of choice) in the form of an online gift voucher (e.g. Amazon voucher). The total annotation cost was approximately 8,150 dollars.

Annotation Process The main annotation process was split into two rounds. In the first round, two annotators were assigned to each tweet-reply pair. In the second round, we collected all tweet-reply pairs on which the two annotators disagreed in their first-choice label during the first round and assigned each of these examples to a third annotator in an effort to obtain a consensus. Thus, each example in the dataset contains a first-choice stance annotation, a confidence score, and a possible second-choice stance annotation from two or three annotators. Annotators were assigned to examples using the EffiARA library Cook et al. (2025). The open-source GATE Teamware tool[6] Wilby et al. (2023) is used for conducting the annotation (see Figure 5.8 in the Appendix for a screenshot of the annotation interface).

5.3.4 Dataset Overview

Label Aggregation

Since each annotator provides multiple signals (i.e., first-choice label, confidence and possible second-choice label), the annotations can be transformed and aggregated in different ways for use in model evaluation. Therefore, we explore and implement eight aggregation methods, including four hard-label-based (e.g., majority voting with or without considering second-choice labels) and four soft-label-based (e.g, Bayesian soft voting (Wu et al. 2023*a*)) methods. We then compute the cosine agreement between the aggregated labels obtained using each pair of methods. Full details and agreement scores are provided in the Appendix. The results suggest that overall the agreement between aggregation methods is high, with the lowest agreement being 0.863. Accordingly, we adopt the conventional *majority-voted first-choice label* as the primary aggregation label for model evaluation in this paper. We also conduct evaluations incorporating the second-choice label to explore the relationship between model predictions and annotator uncertainty.

Dataset Statistics The summary statistics of SCRum-9 are presented in Figure 5.2, where the labels are determined with majority-voting over the first-choice stances as we discussed above. After removing examples which the majority of annotators labelled as *Only refers to image/video*, SCRum-9 contains 7,516 annotated tweet-reply pairs in total. The number of tweets per language ranges from 446 (in Russian) to 1,218 (in Spanish), with 835 tweets per language on average. The data size is comparable to existing rumour stance classification test sets, such as 1,049 and 1,872 in RumourEval 2017 and 2019 official test sets

[5] We did not set a hard agreement threshold, instead evaluating each individual annotator in the context of the others, per language.

[6] `https://annotate.gate.ac.uk/`

respectively. Similar with previous rumour stance classification datasets, the stance class distribution in SCRum-9 is imbalanced, reflecting the natural distribution of stance observed in online rumour-related discussions. The *confirming*, *rejecting*, and *commenting* classes are each relatively evenly represented, accounting for roughly 30% of the dataset, although this proportion varies across languages. The *questioning* class is less represented in SCRum-9, while it normally exhibits the lowest linguistic diversity among the four classes (Li & Scarton 2024).

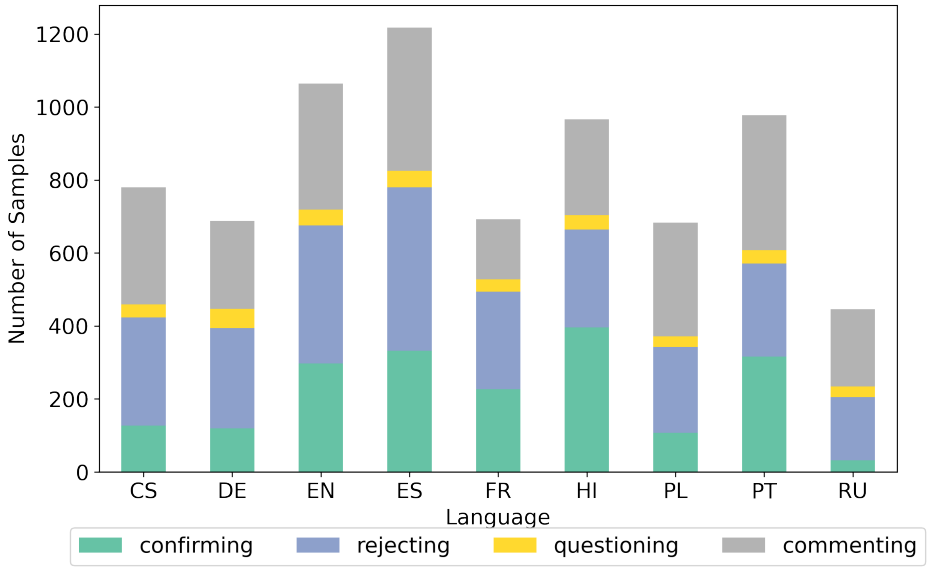


Figure 5.2: *Statistics for SCRum-9 with labels determined with majority-voting over the first-choice stance labels.*

Inter-Annotator Agreement

We calculate the following two kinds of inter-annotator agreement scores:

- **Cosine-based agreement** (Dumitrache et al. 2018): We represent the first-choice label, confidence level and second-choice label of each annotator with vectors, and calculate the cosine agreement scores. The vector encodes the first- and second-choice labels weighted with the confidence scores, followed by normalisation. Details of the vector calculation refer to the *MVC2 (Majority Vote with Confidence and Second-choice)* aggregation method in the Appendix.
- **Percentage-based agreement**: We also report the percentage agreement over the first-choice labels to enable direct comparison with current rumour stance classification datasets (i.e., RumourEval datasets).

As presented in Table 5.2, SCRum-9 achieves an average percentage agreement score of 0.55 (ranging from 0.40 on German to 0.62 on French), comparable to current rumour stance classification datasets, where the agreement score on reply tweets reaches 0.62 in RumourEval

Lang	CS	DE	EN	ES	FR	HI	PL	PT	RU	All
cosine-1	0.55	0.38	0.48	0.51	0.52	0.4	0.47	0.52	0.48	0.49
cosine-2	0.65	0.55	0.62	0.60	0.65	0.47	0.61	0.59	0.59	0.60
percent	0.61	0.40	0.59	0.57	0.62	0.52	0.54	0.58	0.55	0.55

Table 5.2: *Average cosine (`cosine`) and percentage (`percent`) agreement scores across annotators for each language. We report cosine agreement computed using the first-choice label only (`cosine-1`) as well as using the first-choice, second-choice and confidence score (`cosine-2`).*

2017. Furthermore, when incorporating first-choice and second-choice labels along with confidence scores, the cosine-based agreement increases substantially across languages, highlighting the subjective and uncertain nature of this task.

We find that out of 18,280 total annotations, 11,744 (64%) are given a confidence score of 4 or lower, and 4,707 (26%) are assigned a secondary label. Annotator disagreement is most pronounced between the *confirming* and *commenting* categories. We also observe similar pattern on annotators' second-choice labels (as shown in Figure 5.9 in the Appendix). In most of ambiguous cases, annotators select *confirming* or *rejecting* as the first-choice label while considering *commenting* as a possible alternative. The next most frequent pattern is the reverse: annotators select *commenting* as the first-choice label but also consider *confirming* or *rejecting* as plausible options.

5.4 Experiments

In real-world applications, two primary approaches are commonly adopted to perform multilingual classification tasks: applying ICL with multilingual LLMs, or fine-tuning multilingual MLMs on English or multilingual data (Li, Zhao & Scarton 2025, Razumovskaia et al. 2025). In this section, we compare these two approaches across the nine languages in the SCRum-9 dataset. Our experimental setup and analysis focus on: (1) disparities in zero-shot baseline ICL performance across languages and potential strategies to further improve ICL performance, especially on non-English; and (2) the effectiveness of fine-tuning small MLMs with English, translated multilingual data and LLM-generated synthetic multilingual data.

5.4.1 Evaluation Settings

In-Context Learning with LLMs

We evaluate five open source instruction-tuned multilingual LLMs: Qwen2.5[7], Mistral-v0.3[8], Gemma2[9], DeepSeek[10] and Llama3.2[11]. Due to computational constraints, their medium-sized variants are considered.

- **Zero-Shot ICL**: We directly prompt the LLMs by describing the task with natural

[7] `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`
[8] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`
[9] `https://huggingface.co/google/gemma-2-9b-it`
[10] `https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat`
[11] `https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`

language instructions in English following prior work (Zhang, Li, Hauer, Shi & Kondrak 2023). We explore the following three zero-shot ICL settings:

  - `Baseline`: Input text is in the original target languages.
  - `translate-input`: If target-language is non-English, we machine-translate the input text into English. In this study, Google translate[12] is used for machine translation.
  - `align example`: Inspired by recent findings on ICL with extremely low-resource languages (Li, Zhao & Scarton 2025), for languages other than English, we provide the input text in target languages, along with four unlabelled examples in English and their target-language translations. The examples are randomly sampled from the RumourEval 2019 training set, which is the current largest English rumour stance classification dataset.

- **Few-Shot ICL**: We prompt the LLMs with 4-shot demonstration examples. The demonstrations are randomly sampled from each stance category in the RumourEval 2019 training data. We provide the demonstrations in the following three settings:

  - `demo-en`: Demonstration examples are in English.
  - `demo-translate`: If the target input is not in English, we machine-translate the demonstration examples into the target languages.
  - `demo-align`: We provide demonstration examples in both English and their machine translations in the target languages when the input is not in English. Note that this setting is different from `align-example` in zero-shot ICL, whose alignment examples are not provided with the stance labels.

More specifics, including the prompt template and decoding strategy, are provided in the Appendix.

Fine-Tune Multilingual MLMs We evaluate two different MLMs: XLM-R (Conneau et al. 2020) and XLM-T (Barbieri et al. 2022). The models share the same architecture, pretrained with multilingual corpora covering all the languages in SCRum-9. XLM-T is intentionally pre-trained with multilingual Twitter data, while XLM-R is pretrained with filtered CommonCrawl data (Barbieri et al. 2022). We experiment with the following three settings. The hyper-parameter tuning and training details are included in the Appendix:

- `train-en`: Fine-tuning with English data, i.e., the RumourEval 2019 training set.
- `train-translate`: Fine-tuning with machine-translated multilingual data. We machine-translate the English RumourEval 2019 into the eight other SCRum-9 languages.
- `train-synthetic`: Training with multilingual data synthetically generated by multilingual LLMs. We generate the synthetic data with the five LLMs, respectively. Specifically, to ensure each language contains the same number of replies and rumours, we

[12] `https://translate.google.com/`

randomly sample 10 rumours for each language from SCRum-9. For each source tweet, we prompt the LLM to generate N different reply tweets for each stance. The stance defined in each prompt is then used as the labels during fine-tuning. Under the consideration of data efficiency, we opt to generate ten replies per stance per source tweet, resulting in 400 tweets per language and 3,600 multilingual tweets in total[13]. Note that the RumourEval training data we use in the above two approaches contains 6,702 tweets. More details of the prompt template is provided in the Appendix.

Evaluation Metrics Our primary evaluation metric is weighted F_2 score ($wF2$)[14] (Scarton et al. 2020*a*), which rewards models with high performance on the *confirming* and *rejecting* classes, being more adequate to rumour stance classification.

5.5 Results and Discussions

5.5.1 Baseline Zero-Shot ICL Performance and Inconsistency Across Languages

The LLMs exhibit varying performance and cross-lingual consistency when evaluated with baseline zero-shot ICL on SCRum-9. **Most of the LLMs achieve their best performance on English, except for Gemma**. We present the average and standard deviation of the $wF2$ scores across the nine languages for each LLM in Figure 5.3 (detailed results are in the Appendix). LLMs shown in the *bottom-right* area (i.e., high mean with low standard deviation) achieve strong and relatively consistent performance across languages, which is a more desirable outcome. Performances of models on *bottom-left* (i.e., low mean and low standard deviation) are consistently weak, while models on *top-right* (i.e., high mean, high standard deviation) are strong overall but uneven across languages.

The results in Figure 5.3 show that Gemma demonstrates relatively strong and consistent performance across all languages, with the highest $wF2$ of 0.59 on Spanish and the lowest $wF2$ of 0.48 on Hindi, indicating relatively robust cross-lingual generalisation on rumour stance classification. In contrast, Mistral achieves second-best performance on English ($wF2$ as 0.5) among the five LLMs, but its performance on German and Hindi is only 0.33, suggesting substantial variability across languages. Deepseek achieves overall moderate performances with high disparity, while Llama and Qwen show low performances across all languages, suggesting their limited zero-shot ICL capability for this classification task.

5.5.2 ICL for Non-English Rumour Stance Classification: Translation vs. Language Alignment

As discussed above, LLMs show performance disparity across languages, typically performing better in English. Therefore, we analyse which approaches are effective to improve the ICL performance for the eight non-English languages in SCRum-9. We present the results on Qwen, as an example, in Figure 5.4. Full results of other LLMs can be found in the Appendix.

[13]The number of tweets is slightly lower than 3,600 for certain LLMs since they sometimes refuse to generate replies to misinformation.

[14]$wF2$ gives different weights for each stance: $deny = support = 0.40$, $query = 0.15$ and $comment = 0.05$.

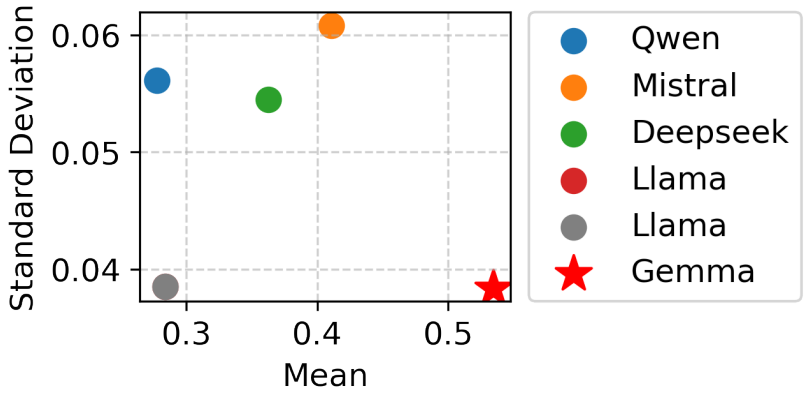


Figure 5.3: *Mean and standard deviation of zero-shot baseline ICL performance (wF2) on SCRum-9 with different LLMs. LLMs on bottom-right with high mean and low standard deviation exhibit good and relatively consistent performance across languages.*

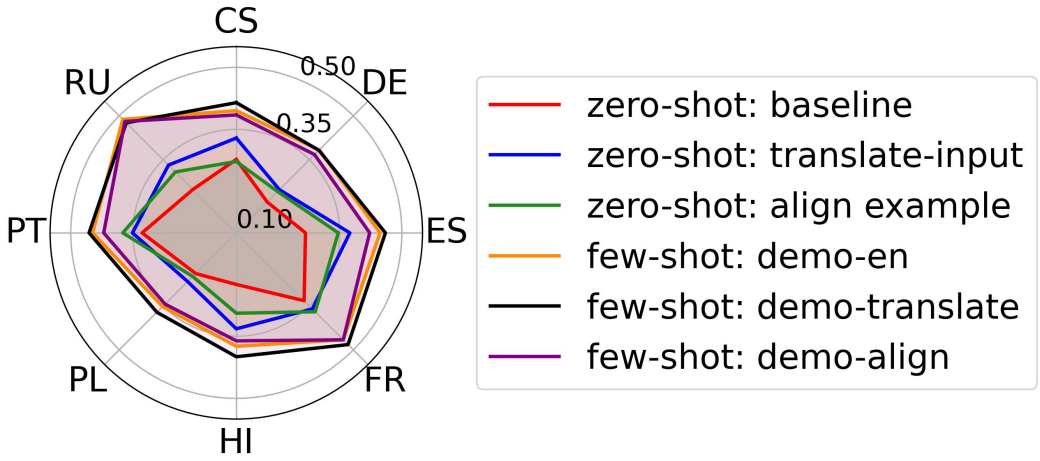


Figure 5.4: *Comparison between ICL performances (wF2) across the eight non-English languages with Qwen.*

Zero-Shot ICL **Both translating the target input into English and providing unlabelled translation pairs as language alignment signals** generally enhance performance over the baseline zero-shot ICL across all the LLMs and languages. However, we observe that the language alignment strategy might slightly degrade performance in certain cases (e.g., on German with Mistral), consistent with previous findings on extremely low-resources languages (Li, Zhao & Scarton 2025). Language alignment also shows no benefit for Llama, which we attribute not only to cross-lingual challenges but also to Llama's overall weak performance on this task, as indicated in Figure 5.3.

Furthermore, **for most languages, directly translating the input into English yields stronger improvements**, except for Gemma for which including the language-alignment signal in the prompt consistently produces better results than translation across

all the languages.

Few-Shot ICL Regardless of whether the demonstration examples are in English or translated into the target non-English, and whether additional language alignment is provided, few-shot ICL generally outperforms zero-shot ICL. This improvement is particularly notable for LLMs such as Llama and Qwen that perform poorly in the zero-shot setting. For example, as shown in Figure 5.4, Qwen shows a clear gap between zero-shot and few-shot ICL performances across all the languages. Also, **in general, providing demonstration examples directly in the target non-English language yields greater gains**. However, Gemma benefits more when using English-only demonstrations for half of the eight non-English languages.

Key Findings We summarise our two key findings for multilingual rumour stance classification with ICL:

- Since the non-English languages in SCRum-9 are relatively high-resource and well supported by machine translation tools (e.g., Google Translate, which we use in this study), **machine translation generally serves as an effective method for improving ICL performance in these languages**.
- Few-shot ICL, even with English demonstration examples, outperforms zero-shot ICL (regardless of whether translation or language alignment strategies are applied), suggesting that **multilingual LLMs are capable of effectively transferring knowledge across languages in ICL, especially the relatively high-resource languages for rumour stance classification**.

5.5.3 Cross-Lingual and Multilingual Fine-Tuning MLMs vs. Prompting LLMs

Although ICL demonstrates promising performance across languages as we discussed above, it is usually computationally more expensive than fine-tuning and then inferencing with smaller multilingual MLMs. We compare the baseline zero-shot ICL and best ICL performance after improvement of Gemma and Llama (i.e., the best and worst LLMs in baseline ICL as shown in Figure 5.3) against XLM-R fine-tuned with either English RumourEval data or its machine-translated multilingual version. The results on XLM-T are similar, which can be found in the Appendix.

As shown in Figure 5.5, we observe **a notable gap between Gemma's ICL performance and that of fine-tuned XLM-R across all the languages**. The performance of fine-tuned XLM-R is comparable to Llama's baseline zero-shot ICL; however, providing only four demonstration examples in the prompt for Llama (e.g., Llama(best) in 5.5) elevates Llama's performance well beyond that of the fine-tuned XLM-R, resulting in a substantial performance gap.

5.5.4 Effectiveness of Multilingual Synthetic Data

In practice, obtaining multilingual training data is both costly and often impractical, as data collection and human annotation require substantial resources. Consequently, a common

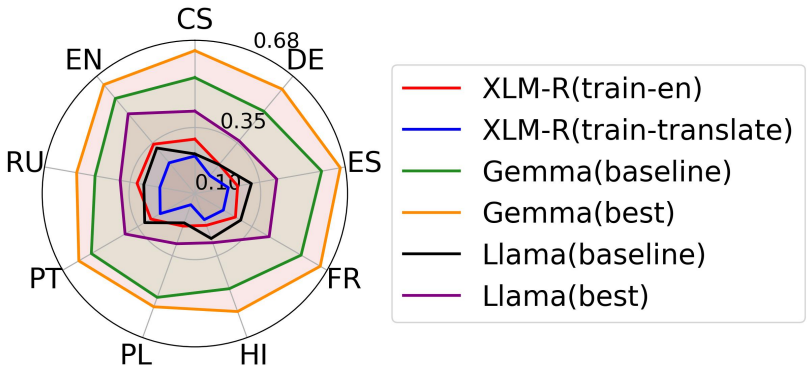


Figure 5.5: *Performance (wF2) comparison across languages between (1) XLM-R fine-tuned with English; (2) XLM-R fine-tuned with translated multilingual data; (3) Baseline zero-shot ICL performances of Gemma and Llama; and (4) Best ICL performances of Gemma and Llama.*

approach is cross-lingual transfer, where multilingual MLM is fine-tuned on English data. However, as shown previously, this produces suboptimal performance, substantially lagging behind ICL. Meanwhile, deploying LLMs for inference in rumour stance classification is not always feasible due to computational and financial constraints. A promising alternative, when direct LLM inference is not feasible, is to leverage LLMs to generate multilingual synthetic data, which can subsequently be used to fine-tune MLMs for deployment.

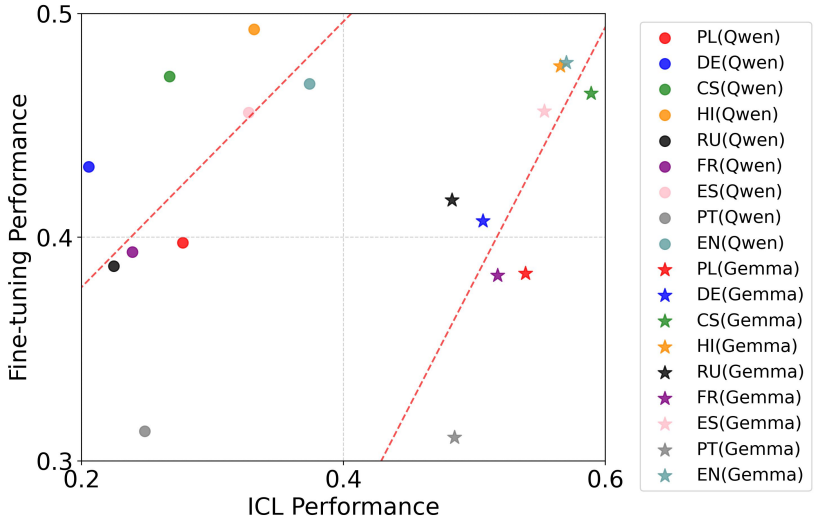
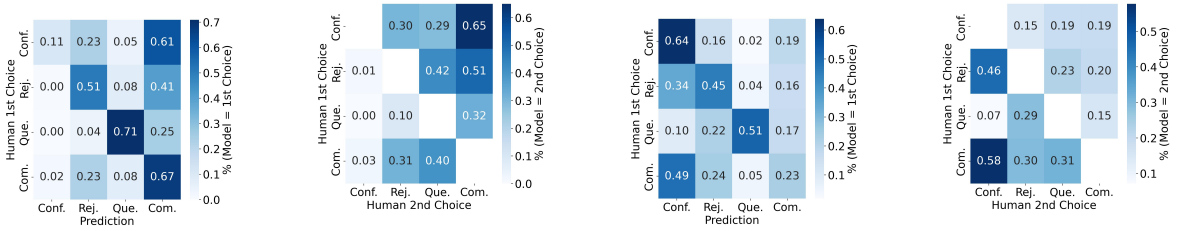


Figure 5.6: *Comparison of Gemma/Qwen baseline zero-shot ICL performance with XLM-R fine-tuned on their generated synthetic data. Gemma is represented by the stars, and Qwen is represented by the circles.*

Overall Performance We find that **fine-tuning MLMs on LLM-generated multilingual synthetic data consistently outperforms training on real-world English data**

or machine-translated multilingual data, despite using nearly almost 3,000 fewer training examples. Also, such fine-tuning achieves performance comparable to, or even exceeding, the zero-shot performance of the same LLMs that produce the synthetic data. In particular, for Qwen, Llama, and DeepSeek – the three LLMs with the weakest baseline zero-shot performance on average (Figure 5.3) – fine-tuning MLMs on their synthetic multilingual data yields significant better performance than ICL. However, for Mistral and Gemma that exhibit the strongest zero-shot baselines on average, fine-tuning with synthetic data might still underperform zero-shot ICL, especially for Gemma. Nonetheless, the performance remains comparable. Full results can be found in the Appendix.



(a) *Baseline Zero-Shot ICL (1st Choice)*

(b) *Baseline Zero-Shot ICL (2nd Choice)*

(c) *XLM-R Train-Synthetic (1st Choice)*

(d) *XLM-R Train-Synthetic (2nd Choice)*

Figure 5.7: *Confusion matrices of (1) baseline zero-shot ICL performance with Deepseek; and (2) XLM-R performance when fine-tuned with sythetic data generated by Deepseek. In Sub-figures (a) and (c), each entry (i, j) in row i column j represents the proportion of tweets (1st choice stance = i) that is classified as stance j by the model. In Sub-figures (b) and (d), each entry (i, j) denotes the proportion of tweets (1st choice stance = i, 2nd choice stance = j) that is classified as stance j by the model.*

Can ICL performance predict the usefulness of synthetic data? To investigate whether LLMs with stronger baseline ICL performance result in more effective synthetic data, we analyse the relationship between LLM's ICL performance and the performance of MLMs fine-tuned on synthetic data generated by this LLM.

Within each LLM across languages, we observe that **stronger ICL performance in a language tends to indicate more effective synthetic data generation** that leads to higher fine-tuning performance (as shown in Figure 5.6). Notably, **poor ICL performance does not necessarily result in poor fine-tuning outcomes**, which is consistent across all the five LLMs we study. For example, although Qwen's baseline ICL on German is the lowest among all the languages in SCRum-9 ($wF2 = 0.21$ as shown in Figure 5.6), fine-tuning XLM-R on its synthetic data yields a $wF2$ score as 0.43, higher than other languages (e.g., Portuguese, Russian, French and Polish) where Qwen achieves better ICL performance but produces less effective synthetic data.

Furthermore, we find that the effectiveness of synthetic data also depends on the choice of MLM. For instance, Qwen's synthetic data leads to the highest average performance on XLM-R, while Gemma's generated data performs best on XLM-T.

Overall, practitioners could consider generating multilingual synthetic data for rumour stance classification in relatively high-resource languages with LLMs that exhibit strong ICL

performance. **Fine-tuning MLMs on such data has the potential to match, and in some cases even exceed, the original ICL performance**.

5.5.5 Model Prediction vs. Human Uncertainty

For ambiguous cases when human annotators provide a second-choice label, if the model predictions differ from the aggregated first-choice label but match the aggregated second-choice label, we assume the model still captures a plausible interpretation of the instance that aligns with human uncertainty. To capture this, we re-calculate our evaluation metrics by considering a prediction correct if it matches either the first-choice or second-choice label.

We observe a substantial increase in $wF2$ scores across all the models, approaches and languages (full results see the Appendix). For example, $wF2$ score improves from 0.39 to 0.5 on Polish with Mistral in baseline zero-shot setting, when considering the second-choice human label. It suggests that, **in those ambiguous cases, the models are capable of generating predictions that reflect plausible alternative interpretations consistent with human uncertainty, not merely incorrect in a strict sense.**

As discussed before (also see Figure 5.9), most ambiguous cases involve annotator hesitation between *confirming* and *commenting* or between *rejecting* and *commenting*. We find that models handle these cases differently. Generally, MLMs trained with synthetic data exhibits different patterns than the other approaches we study in this paper. We present two examples related to Deepseek in Figure 5.7. The results suggest that in those ambiguous cases, Deepseek in baseline zero-shot ICL setting tends to predict *commenting* (Figure 5.7b), whereas XLM-R fine-tuned with Deepseek-generated synthetic data more often predicts *confirming* or *rejecting* (Figure 5.7d). This pattern is also reflected in their overall performance (Figure 5.7a and 5.7c), where the former tends to over-predict *commenting* and the latter tends to over-predict *confirming* or *rejecting*.

5.6 Conclusion

We introduce SCRum-9, the largest multilingual rumour stance classification benchmark to date, covering 7,516 instances in nine languages. The dataset incorporates annotator uncertainty through a second-choice label design, also enabling studies of uncertainty and subjectivity in stance classification. Our experiments reveal substantial performance disparities across languages in ICL with LLMs, while outperforming or matching MLMs fine-tuned with English or machine-translated multilingual data. However, MLMs fine-tuned on synthetic multilingual data generated by LLMs yields competitive results, in some cases surpassing ICL with the same LLM, while being more computationally efficient. We believe our work will serve as a valuable resource for advancing multilingual stance classification and online rumour analysis.

Limitations

While we go beyond previous work by assigning multiple annotators per example, budget and annotator availability limited us to only two or three annotators. This is less than ideal for obtaining a representative sample, so future work might want to obtain annotations from

additional annotators. Nevertheless, we highlight the importance of our novel dataset, given the complexity of the task.

Our model evaluation did not take full advantage of the label aggregation methods, to our knowledge, there are no accepted methods for prompting LLMs to predict a distribution over labels. Future work ought to determine the extent to which LLMs can predict such distributions, and evaluate them against the various label aggregations described above. Also, although we analyse the effectiveness of LLM-generated synthetic data, we did not perform data analysis or human analysis to evaluate the multilingual data quality (e.g., lexical diversity or whether the generated replies truly express the stance required in the prompt) mainly due to lack of native speakers in the nine target languages. Although our experimental results have demonstrate their effectiveness, future work could explore how it correlates with the characteristics of the generated multilingual data.

While our evaluation focused on stance classification, because SCRum-9 links each source tweet to fact-checked claims, it is also possible to perform claim verification, which we did not evaluate here. Future work is required to build and evaluate claim verification models on our dataset.

Finally, although we make efforts to cover nine languages, the languages in this study are relatively high-resource languages, and future work should consider establishing datasets for low-resource languages, and benefiting more under-represented communities.

Ethical Considerations

We anonymised the tweets by removing user mentions. This was done primarily for anonymity during annotation. However, because the tweets are indexed in our dataset by tweet ID, it is simple to deanonymise them by looking them up on X directly or using the API. Keying by the tweet IDs is necessary for our dataset to be used by other researchers, as the X developer agreement prohibits redistribution of tweet texts, instead requiring users to use the X API to hydrate tweets keyed by their ID. The dataset will be released with a CC BY-NC-SA 4.0 license.[15]

We received ethics approval by our institution (omitted due to double-blind constraints) to conduct this research. We consider social media data as a type of personal data where consent is not feasible to obtain. Nevertheless, following legislation, we are allowed to collect data which will result in research of public interest. Since the collected tweets are flagged by fact-checkers as potential source of disinformation, their collection is justified. In addition, we also received ethical approval for the annotation experiments, complying with the best practices for information sheet and consent forms creation as well as data anonymisation prior to annotation.

All data is securely stored in our encrypted servers. The data shared with annotators is done via annotation tool, that also stores all data in an encrypted servers. Annotators data (e.g. e-mail) are stored separately from the annotations and researchers only have access to annotators' IDs and annotators (i.e. no access to personal information is available). The dataset is made available for research without any identification to the annotators.

Finally, we advise all users of our dataset that because the tweets were sourced from X they contain potentially offensive content.

[15] `https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en`

5.7 Appendix

5.7.1 Data Collection and Filtering

We provide URLs of each fact-checking source used during data collection in Table 5.3. Additional details and scripts used to obtain data from DBKF and the fact-checking websites will be made available with the publicly available source code. The statistics of the source tweets, replies, and fact-checked claims that were collected, filtered, and annotated can be found in Table 5.4.

CS	`dbkf.ontotext.com`; `napravoumiru.afp.com`
	`demagog.cz`
DE	`dbkf.ontotext.com`
EN	`factcheck.afp.com`; `dbkf.ontotext.com`
ES	`factual.afp.com`; `dbkf.ontotext.com`
FR	`factuel.afp.com`; `dbkf.ontotext.com`
PL	`sprawdzam.afp.com`; `demagog.org.pl`
	`dbkf.ontotext.com`
PT	`checamos.afp.com`; `dbkf.ontotext.com`
RU	`dbkf.ontotext.com`

Table 5.3: *Source fact-check websites used for sourcing the X links.*

We use the `bertopic` Python library to perform the topic modelling.[16] Specifically, we use KeyBertInspired as the representation model and a UMAP model 15 neighbours, 5 components, and cosine distance to perform dimensionality reduction. We then perform a manual review of the resulting topics, keeping only those that we deemed coherent, and merging topics that were closely related. Of the resulting 63 topics, the five most common are Russia-Ukraine (774 source tweets), COVID-19 (440), US Elections (301), Israel-Palestine (201), and Natural Disasters (133), all of which are represented by the six languages (DE, EN, ES, FR, PL, PT) for which we used topic model filtering. We also report the five most common topics per language in Table 5.5.

5.7.2 Annotation Guideline and Interface

A screenshot of the annotation interface is presented in Figure 5.8. The full text of the annotation guidelines is reproduced below:

1. **Stance Annotation:** After reading the source and reply tweets thoroughly, you will need to decide the stance of the reply tweet towards the source tweet based on the following definitions. Please ONLY rely on the given tweets and refrain from using additional resources in this step.
    - Confirming: The author of the response supports the source tweet.
    - Rejecting: The author of the response disagrees with the source tweet.

[16] `https://maartengr.github.io/BERTopic/index.html`

Lang	Collected			Filtered			Annotated		
	Sources	Replies	Claims	Sources	Replies	Claims	Sources	Replies	Claims
CS	89	918	128	24	780	14	24	780	14
DE	358	4,548	528	49	1,500	56	24	705	19
EN	15,734	215,152	19,529	1,402	1,500	1,565	1,026	1,100	1,169
ES	7,894	70,333	11,518	491	1,500	530	414	1,260	419
FR	1,477	20,159	2,864	224	1,500	286	109	700	122
HI	1,668	13,601	2,130	320	1,500	342	220	1,000	227
PL	258	4,105	404	42	1,500	45	23	700	224
PT	2,485	22,473	3,845	197	1,500	204	124	1,000	133
RU	59	831	95	20	495	31	19	450	29
All	30,022	352,120	41,041	2,769	11,775	3,073	1,983	7,695	2,156

Table 5.4: *Statistics of the source tweets, replies, and fact-checked claims that were collected, filtered, and annotated.*

DE	Russia-Ukraine, COVID-19, US Elections, Climate Change, Israel-Palestine
EN	Russia-Ukraine, COVID-19, US Elections, Israel-Palestine, Natural Disasters
ES	Russia-Ukraine, COVID-19, Venezuela Dictatorship, US Elections, Natural Disasters
FR	Russia-Ukraine, COVID-19, Climate Change, US Elections, Israel-Palestine
PL	Russia-Ukraine, COVID-19, US tax legislation, The Kashmir Files Film, US Immigration
PT	COVID-19, Brazil President Lula, Russia-Ukraine, US Elections, COVID-19

Table 5.5: *The top five most common topics per language.*

- Questioning: The author of the response asks for additional evidence in relation to the source tweet.
- Commenting: The author of the response makes their own comment without a clear stance towards the source tweet. It includes replies that are unrelated to the source tweet.
- Only refers to image/video: The author of the response makes the comment only/mainly referring to the image or video attached to the source tweet.

Note that sarcastically or humorously supporting the source tweet should be considered as rejecting, while sarcastically or humorously denying source tweet should be considered as confirming. Also, the appearance of question marks does NOT necessarily indicate the questioning stance. For example, rhetorical questions should not be coded as questioning, since the author does not expect an answer.

2. **Confidence Rating:** Please indicate how confident you are about your annotation. The confidence scores range from 1 to 5 and hold the following meaning. You will need to indicate a second-choice stance label if your confidence score is lower than 3, including 3.

   - 5 - extremely confident about the annotation (I'm certain about the annotation without a doubt.)
   - 4 - fairly confident about the annotation (I'm confident about the annotation, but might be in small chance other annotators may label it in a different category)

- 3 - pretty confident about the annotation (I'm pretty sure about the annotation, but might be in high chance other annotators may label it in a different category)
- 2 - not confident about the annotation (I'm not sure about the annotation, it seems it also belongs to other categories, but you can still include this instance as a "silver standard instance" in training)
- 1 - extremely unconfident about the annotation (I'm really unsure about the annotation. It may belong to another category as well, you may wish to discard this instance from the training.)

3. **Second-choice Label:** If your confidence score is lower than 3 including 3, please select an alternative category that the stance of the target tweet may also belong to, except for the following conditions: (1) *Not Applicable*: Your confidence score is larger than 3; (2) *Highly Uncertain*: You are unable to provide a second-choice label due to significant uncertainty about it.

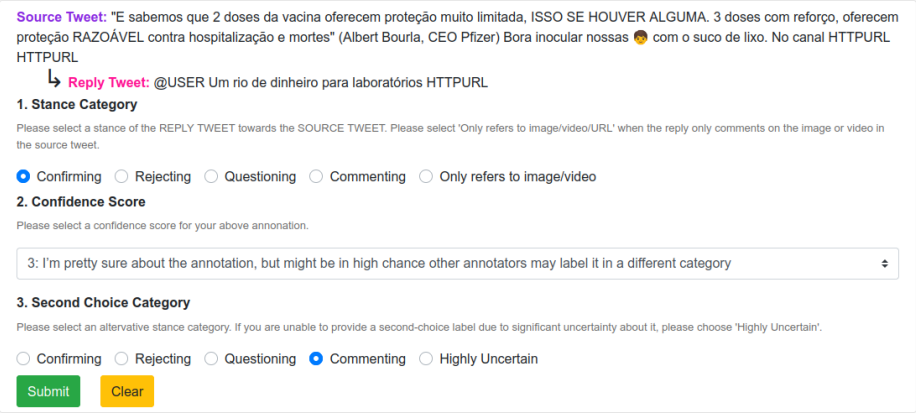


Figure 5.8: *An example of the GATE Teamware annotation interface. The Second Choice Category section only appears when the provided Confidence Score is 3 or lower. The tool also ensures that annotators do not choose the same label for the first and second choice.*

5.7.3 Second-Choice Label Analysis

Figure 5.9 presents the confusion matrix between annotators' first-choice and second-choice labels, aggregated via majority voting. The results show that, in most ambiguous cases, annotators select *confirming* or *rejecting* as the first-choice label while considering *commenting* as a possible alternative. The next most frequent pattern is the reverse: annotators select *commenting* as the first-choice label but also consider *confirming* or *rejecting* as plausible options.

5.7.4 Label Aggregation

Aggregation methods: We categorise the methods into *hard label*-based and *soft label*-based.

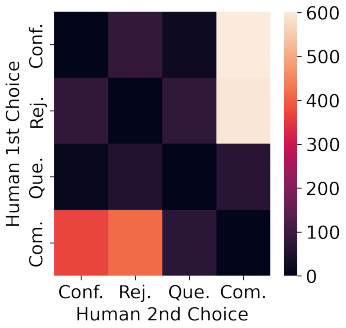


Figure 5.9: *Confusion matrix between annotators' first-choice and second-choice labels. Each entry (i, j) in row i column j denotes the number of tweets whose first-choice label is stance i and second-choice label is stance j. The labels are aggregated through majority-voting.*

1. Hard labels: These methods result in a single label being assigned to each example, i.e., a one-hot vector.

    - **Majority Vote (MV):** We take the most common first-choice label across annotators. Where there is no consensus, we chose a label at random from those annotated.
    - **Majority Vote with Confidence (MVC):** Instead of treating each annotation equally, we weight the first-choice class by the annotator's reported confidence, normalized into the range $[0, 1]$. The class with the greatest confidence-weighted count among the annotators is chosen.
    - **Majority Vote with Confidence and Second-choice (MVC2):** The same as MVC, but we also add in the second-choice label discounted by 2 times the normalised confidence score.

2. Soft labels: These methods result in a categorical distribution over labels.

    - **Soft Vote (SV), SV with Confidence (SVC), and SVC and Second-choice (SVC2):** These three methods work the same as the corresponding hard label methods described above. However, instead of returning a single label, we return the distribution over labels computed by normalising the (weighted) counts into the range $[0, 1]$.
    - **Dawid-Skene (DS):** This method computes a distribution over labels by estimating a probabilistic graphical model of annotation errors in order to discover the underlying true label Dawid & Skene (1979).
    - **Bayesian Soft Vote (BSV):** This method was developed specifically for cases such as ours where annotators provide a confidence score. While each annotator chooses their confidence from the same 1-5 Likert scale, their individual perceptions of what each step on the scale means may differ from one another. To account for this, Bayesian Soft Vote recalibrates each annotator's confidence scores according to their level of agreement with other annotators by estimating a probabilistic graphical model Wu et al. (2023*a*).

For the methods that use the second-choice annotations, if the annotators choose *Highly Uncertain* as their second-choice, we uniformly redistribute the probability mass remaining after the first-choice annotation to all other labels.

Comparison of Label Aggregation Methods We compare the above label aggregation methods by computing the cosine agreement between the aggregated labels obtained using each pair of methods. As shown in Figure 5.10, the overall the agreement between aggregation methods is high, with the lowest agreement being only 0.863. Note that a relatively low agreement (e.g., BSV) is not necessarily negative, as it simply shows that the method encodes information differently than the others, and there is not necessarily one "best" way to use the annotations.

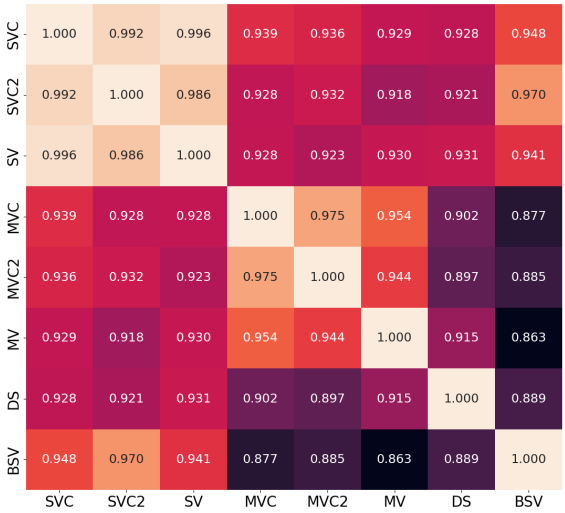


Figure 5.10: *Cosine agreements among label aggregation methods, ordered from top-to-bottom according to the average overall agreement with the other methods.*

5.7.5 Experimental Setups

All the experiments were run on a single Nvidia A100 GPU with 40GB of VRAM.

Synthetic Data Generation We use the following prompt: *Given a source tweet, generate 10 different replies in {language_name} that {stance} the source tweet. Number each reply and put it on a new line. The source tweet is:{source_tweet}.* We use sampling-based decoding, with temperature as 0.3.

Zero-shot ICL We use greedy search in decoding to ensure reproducibility. In the following prompt templates, `target_input` denotes "*Source tweet: {source_tweet}. Reply tweet: {reply_tweet}*", and `task_instruction` is "*Source tweet: {source_tweet}. Reply tweet: {reply_tweet}. Determine the stance of the reply tweet towards the source tweet. The possible stance labels are "support", "deny", "query", or "comment". Answer with the stance label first, before*

any explanation. Definitions of the stance labels follow. "support": the reply tweet agrees with the source tweet. "deny" the reply disagrees with the source tweet. "query": the reply tweet asks for more information regarding the source tweet. "comment": the reply tweet does not take a clear stance towards the source tweet."

- **Baseline**: *{target_input}{task_instruction}* with target inputs in original languages.
- **translate-input**: *{target_input}{task_instruction}* with target inputs translated into English.
- **align example**: we use the following prompt: *Use the following pairs of {language name} texts and their English translation to help you understand {language name}. Example {num_i}: {language name}: {tweet}. English: {English tweet}. Now based on your understanding, answer the question below. {target_input}{task_instruction}.*

Few-Shot ICL We use greedy search in decoding to ensure reproducibility.

- **demo-en**: We use the following prompt: *{task_instruction} Example {num_i}: Source tweet: {source_tweet}. Reply tweet: {reply_tweet}. Stance is: {stance}. Now complete the following example and answer with the stance label first before any explanation. {target_input} Stance is*
- *demo-translate*: We use the same prompt template as above, while the examples are translated into target non-English.
- *demo-align*: *{task_instruction} Example {num_i}: Source tweet: {source_tweet}. {language name} translation is: source tweet translation. Reply tweet: {reply_tweet}. {language name} translation is: {reply tweet translation}. Stance is: {stance}. Now complete the following example and answer with the stance label first before any explanation. {target_input} Stance is*

Fine-Tuning MLMs We use AdamW optimizer and search batch size from [16, 32] and learning rate from [1e-5,5e-5,1e-6,2e-6]. The optimal hyper-parameters are determined based on $wF2$ on RumourEval 2019 validation set. The number of training epochs is set as 5.

5.7.6 Experimental Results

		CS	DE	ES	FR	HI	PL	PT	RU	EN
XLM-R(Qwen)	1st	0.3976	0.4315	0.4718	0.4929	0.3870	0.3933	0.4556	0.3133	0.4685
	2nd	0.4676	0.5198	0.5210	0.5700	0.4327	0.4840	0.5281	0.4321	0.5572
XLM-R(Deepseek)	1st	0.4769	0.4962	0.5181	0.5156	0.4490	0.4197	0.4907	0.3342	0.5426
	2nd	0.5551	0.5843	0.5724	0.6096	0.4905	0.5533	0.5645	0.4425	0.6404
XLM-R(Llama)	1st	0.4282	0.4590	0.4147	0.4050	0.3350	0.3764	0.3794	0.3500	0.4287
	2nd	0.4895	0.5304	0.4599	0.4790	0.3838	0.4552	0.4485	0.4378	0.5154
XLM-R(Gemma)	1st	0.3838	0.4072	0.4642	0.4765	0.4165	0.3829	0.4564	0.3105	0.4780
	2nd	0.4403	0.4985	0.5152	0.5460	0.4665	0.4865	0.5167	0.4149	0.5621
XLM-R(Mistral)	1st	0.4120	0.4510	0.4808	0.4903	0.4178	0.3658	0.4443	0.3476	0.4658
	2nd	0.4777	0.5397	0.5357	0.5787	0.4733	0.4879	0.5217	0.4778	0.5662
XLM-T(Qwen)	1st	0.4338	0.4528	0.5371	0.5010	0.4666	0.3899	0.5121	0.3797	0.4973
	2nd	0.4882	0.5244	0.5724	0.5630	0.5106	0.4803	0.5709	0.4884	0.5830
XLM-T(Deepseek)	1st	0.4343	0.4910	0.4629	0.4855	0.3996	0.3469	0.4357	0.2716	0.4957
	2nd	0.5208	0.5996	0.5223	0.5791	0.4574	0.4896	0.5183	0.3758	0.5861
XLM-T(Llama)	1st	0.4108	0.4559	0.4869	0.4807	0.3256	0.4255	0.4712	0.4223	0.4357
	2nd	0.4737	0.5251	0.5432	0.5566	0.3776	0.5460	0.5298	0.5559	0.5298
XLM-T(Gemma)	1st	0.4604	0.5174	0.5454	0.5350	0.4660	0.4737	0.5295	0.4441	0.5596
	2nd	0.5106	0.5931	0.5815	0.5931	0.4916	0.5648	0.5925	0.5497	0.6312
XLM-T(Mistral)	1st	0.3564	0.3653	0.4455	0.4329	0.4012	0.2944	0.4179	0.2264	0.4414
	2nd	0.4249	0.4770	0.4981	0.5224	0.4493	0.4015	0.4873	0.3257	0.5417

Table 5.6: *Fine-tuning MLM with synthetic data, evaluated on the first choice and second choice labels.*

		CS	DE	ES	FR	HI	PL	PT	RU	EN
baseline(Qwen)	1st	0.2773	0.2056	0.2672	0.3318	0.2247	0.2390	0.3278	0.2483	0.3741
	2nd	0.3478	0.3166	0.3053	0.4280	0.2653	0.3830	0.4103	0.3215	0.4803
translate-input (Qwen)	1st	0.3292	0.2475	0.3744	0.3600	0.3319	0.2664	0.3509	0.3316	-
	2nd	0.4160	0.3643	0.4214	0.4650	0.4012	0.3487	0.4164	0.4032	-
align example (Qwen)	1st	0.2725	0.2413	0.3465	0.3697	0.2943	0.2491	0.3743	0.3082	-
	2nd	0.3417	0.3606	0.3910	0.4860	0.3517	0.3794	0.4433	0.3946	-
baseline(Mistral)	1st	0.3847	0.3324	0.4376	0.4378	0.3316	0.3618	0.4663	0.4426	0.5032
	2nd	0.4695	0.4458	0.4970	0.5489	0.3906	0.4605	0.5466	0.5470	0.6208
translate-input (Mistral)	1st	0.4289	0.3952	0.4843	0.4547	0.4360	0.3898	0.4631	0.4472	-
	2nd	0.5114	0.5065	0.5425	0.5706	0.5023	0.4882	0.5416	0.5532	-
align example (Mistral)	1st	0.4310	0.3644	0.4308	0.4603	0.3722	0.3894	0.4473	0.4445	-
	2nd	0.5079	0.4650	0.4879	0.5586	0.4276	0.5007	0.5367	0.5552	-
baseline(Gemma)	1st	0.5391	0.5066	0.5893	0.5657	0.4830	0.5178	0.5535	0.4849	0.5702
	2nd	0.6120	0.6431	0.6491	0.6590	0.5351	0.6346	0.6190	0.5938	0.6758
translate-input (Gemma)	1st	0.6034	0.5298	0.6157	0.6016	0.5508	0.5099	0.5908	0.5368	-
	2nd	0.6540	0.6571	0.6699	0.6912	0.6027	0.6067	0.6580	0.6583	-
align example (Gemma)	1st	0.6412	0.6158	0.6610	0.6504	0.5754	0.5558	0.6092	0.5564	-
	2nd	0.6986	0.7124	0.7069	0.7313	0.6226	0.6428	0.6700	0.6682	-
baseline(Deepseek)	1st	0.3391	0.3450	0.4161	0.4100	0.3097	0.2884	0.3910	0.3186	0.4446
	2nd	0.4155	0.4438	0.4624	0.4903	0.3626	0.3966	0.4435	0.4051	0.5486
translate-input (Deepseek)	1st	0.3674	0.3696	0.3886	0.3929	0.3470	0.3339	0.3663	0.3370	-
	2nd	0.4481	0.4557	0.4384	0.4813	0.4019	0.4149	0.4207	0.4093	-
align example (Deepseek)	1st	0.3073	0.3259	0.3904	0.4196	0.2380	0.3184	0.3707	0.3087	-
	2nd	0.3720	0.4173	0.4374	0.4790	0.2793	0.4162	0.4234	0.4106	-
baseline(Llama)	1st	0.2500	0.2434	0.3181	0.3008	0.2810	0.2176	0.3205	0.2971	0.3254
	2nd	0.2985	0.3023	0.3539	0.3499	0.3173	0.2878	0.3664	0.3695	0.3759
translate-input (Llama)	1st	0.3131	0.2651	0.3027	0.2912	0.2864	0.2811	0.3020	0.3290	-
	2nd	0.3595	0.3135	0.3371	0.3392	0.3183	0.3325	0.3423	0.3779	-
align example (Llama)	1st	0.1228	0.1533	0.1880	0.1986	0.1241	0.1384	0.1952	0.1394	-
	2nd	0.1425	0.1896	0.2160	0.2503	0.1448	0.1835	0.2358	0.1802	-

Table 5.7: *Zero-shot ICL performance evaluated on the first choice and second choice labels.*

		CS	DE	ES	FR	HI	PL	PT	RU	EN
demo-en(Qwen)	1st	0.3954	0.3833	0.4475	0.4657	0.3740	0.3520	0.4473	0.4881	0.5190
	2nd	0.4697	0.4824	0.4997	0.5605	0.4270	0.4477	0.5108	0.5969	0.6169
demo-translate(Qwen)	1st	0.4147	0.3823	0.4604	0.4819	0.3995	0.3714	0.4558	0.4761	-
	2nd	0.4868	0.4773	0.5172	0.5744	0.4628	0.4663	0.5125	0.5886	-
demo-align(Qwen)	1st	0.3850	0.3675	0.4223	0.4655	0.3610	0.3435	0.4205	0.4807	-
	2nd	0.4556	0.4597	0.4692	0.5415	0.4187	0.4325	0.4803	0.6028	-
demo-en(Mistral)	1st	0.4893	0.4556	0.5189	0.5054	0.4103	0.3969	0.5003	0.5405	0.5742
	2nd	0.5572	0.5635	0.5759	0.6118	0.4654	0.5014	0.5679	0.6508	0.6742
demo-translate(Mistral)	1st	0.4981	0.4530	0.5229	0.5151	0.4364	0.4025	0.5133	0.5274	-
	2nd	0.5693	0.5635	0.5806	0.6220	0.4923	0.5074	0.5850	0.6386	-
demo-align(Mistral)	1st	0.4624	0.4271	0.4906	0.4779	0.3663	0.3752	0.4725	0.5167	-
	2nd	0.5346	0.5225	0.5396	0.5747	0.4205	0.4726	0.5496	0.6073	-
demo-en(Gemma)	1st	0.6138	0.5948	0.6382	0.6260	0.5636	0.4969	0.6069	0.5542	0.6382
	2nd	0.6931	0.6989	0.6897	0.7203	0.6130	0.6020	0.6711	0.6489	0.7418
demo-translate(Gemma)	1st	0.6109	0.5761	0.6416	0.6285	0.5472	0.5047	0.5998	0.5618	-
	2nd	0.6917	0.6921	0.6896	0.7184	0.6007	0.6209	0.6657	0.6545	-
demo-align(Gemma)	1st	0.5587	0.5207	0.5873	0.5858	0.5057	0.4655	0.5646	0.5634	-
	2nd	0.6347	0.6331	0.6367	0.6705	0.5602	0.5686	0.6263	0.6483	-
demo-en(Deepseek)	1st	0.3830	0.4016	0.4589	0.4481	0.3662	0.3649	0.4227	0.4254	0.5036
	2nd	0.4527	0.5049	0.5119	0.5453	0.4127	0.4644	0.4868	0.5031	0.5917
demo-translate(Deepseek)	1st	0.4392	0.4446	0.4846	0.4374	0.4077	0.3950	0.4359	0.4832	-
	2nd	0.5072	0.5308	0.5352	0.5245	0.4524	0.4874	0.4940	0.5514	-
demo-align(Deepseek)	1st	0.3641	0.3650	0.4411	0.3901	0.3637	0.3430	0.4100	0.4038	-
	2nd	0.4351	0.4561	0.4909	0.4696	0.4075	0.4527	0.4733	0.4930	-
demo-en(Llama)	1st	0.4009	0.3579	0.3984	0.4149	0.3156	0.3111	0.3927	0.3896	0.4947
	2nd	0.4679	0.4738	0.4447	0.5137	0.3647	0.4128	0.4666	0.4844	0.5869
demo-translate(Llama)	1st	0.4117	0.3609	0.4156	0.4257	0.2983	0.3025	0.4057	0.3884	-
	2nd	0.4886	0.4809	0.4614	0.5223	0.3484	0.3982	0.4808	0.4824	-
demo-align(Llama)	1st	0.2932	0.2349	0.3209	0.3019	0.2650	0.2173	0.3362	0.3528	-
	2nd	0.3557	0.3246	0.3597	0.3977	0.3094	0.2848	0.4073	0.4413	-

Table 5.8: *Few-shot ICL performance evaluated on the first choice and second choice labels.*

		CS	DE	ES	FR	HI	PL	PT	RU	EN	RumourEval
train-en(XLM-R)	1st	0.3062	0.2435	0.2648	0.2783	0.2280	0.2307	0.2930	0.3231	0.3443	0.4823
	2nd	0.3901	0.3525	0.2991	0.3675	0.2656	0.3284	0.3523	0.4253	0.4188	-
train-translate(XLM-R)	1st	0.2417	0.1895	0.2287	0.2250	0.2051	0.1453	0.2518	0.2354	0.2519	-
	2nd	0.2993	0.2606	0.2637	0.3114	0.2411	0.2742	0.3105	0.3005	0.3409	-
train-en(XLM-T)	1st	0.2557	0.2205	0.2916	0.2900	0.2095	0.2293	0.2934	0.3480	0.3445	0.4616
	2nd	0.3329	0.3112	0.3334	0.3928	0.2503	0.3345	0.3551	0.4626	0.4413	-
train-translate(XLM-T)	1st	0.1290	0.1564	0.1827	0.1958	0.1453	0.1635	0.2439	0.2271	0.2479	-
	2nd	0.1679	0.2266	0.2066	0.2557	0.1755	0.2212	0.2884	0.2770	0.3205	-

Table 5.9: *Fine-tuning MLMs performance evaluated on the first choice and second choice labels. We also provide its performance on RumourEval 2019 test set for reference.*

Chapter 6

Conclusions

This thesis presents novel methods for generalising and adapting rumour stance classification models across domains and languages. In this chapter, we summarise the findings and contributions, and discuss the future research directions.

6.1 Summary of Thesis

Publication I: *Can We Identify Stance Without Target Arguments? A Study for Rumour Stance Classification* included in Chapter 2 presents our novel insights on how to improve model generalisabity, inspired by the task-specific characteristics that distinguish rumour stance classification from generic stance classification. We highlight the existence of the reply posts whose stance can be naturally inferred without knowing the rumour, and demonstrate its unexpected impact on current systems that generalise poorly to *target-dependent* replies. We propose a novel ensemble-based method that employs a cross-attention mechanism and sample-weighted training to improve the reasoning with the target rumours, achieving state-of-the-art performance on the two RumourEval datasets.

Publication II: *Rumour Stance Classification Adaptation Under Domain Shift: a case study of rumours in Ireland* in Chapter 3 presents our study on domain adaptation for rumour stance classification. We first introduce the `ISLES` test set that exhibits substantial domain shift when compared to the current widely adopted RumourEval datasets. We then evaluate the generalisabity of current supervised models under domain shift, observing substantial performance drop. Finally, we propose a new self-training framework integrated with unlabelled real-world data and synthetic data generated by LLMs, demonstrating a remarkable improvement of nearly 35 points in $wF2$ on the target domain after adaptation. The method is data-efficient, without requiring access to human labelled data in both source and target domains.

Publication III: *Label Set Optimization via Activation Distribution Kurtosis for Zero-Shot Classification with Generative Models* in Chapter 4 presents the first systematic study on how label option design in the prompt (i.e., lexical choice, order, and elaboration) affects classification performance in zero-shot ICL. We find that classification performance varies substantially across different label sets on multiple stance and topic classification datasets.

We empirically demonstrate that optimal label sets produce fewer outlier neurons in LLMs' feed-forward networks. Based on this insight, we then propose LOADS, a novel post-hoc method for optimal label set selection in zero-shot ICL, outperforming common strategy that directly adopts the original dataset labels. Our method only requires 100 unlabelled data examples, demonstrating effectiveness across datasets, LLMs and languages, including rumour stance classification in English, French and Portuguese.

Publication IV: *SCRum-9: Multilingual Stance Classification over Rumours on Social Media* in Chapter 5 presents our study on model adaptation to high-resource and medium-resource languages for rumour stance classification. We introduce SCRum-9, the current largest multilingual rumour stance classification datasets in nine languages. We benchmark four LLMs, observing notable performance disparities across languages. Our findings suggest that machine translation (i.e., non-English to English for target input, or English to target non-English for demonstration examples) is an effective and competitive strategy to improve ICL performance on languages with relatively reliable translation quality. Furthermore, MLMs fine-tuned on LLM-generated multilingual data achieve results comparable to or surpassing zero-shot ICL with the same LLM, while requiring far less computational resources. Finally, SCRum-9 also contains information about annotators uncertainty. We find that in ambiguous cases when annotators provide secondary labels, model predictions could align with these alternatives, reflecting human-like uncertainty rather than simple errors.

6.2 Research Questions Discussion

RQ1: What is the role of targets in the generalisation of rumour stance classification? How can models adequately use information from targets? Publication I empirically demonstrates the substantial presence of target-independent cases in rumour stance classification that was overlooked in prior work but critically affects model generalisation to target-dependent cases. We manually categorise the tweets in the RumourEval Twitter test set into target-dependent and target-independent cases, and find that current models generalise poorly to target-dependent tweets for unseen rumours. We propose a novel ensemble-based framework, where a siamese network with cross-attention is adopted to enhance reasoning with the targets and sample-weighted training is utilised to encourage the model to prioritise potential target-dependent cases. Our framework achieves state-of-the-art performance on two rumour stance classification datasets.

RQ2: Can state-of-the-art rumour stance classification models effectively generalise across domains? How can their performance be improved via adaptation to new domains? In Publication II, we introduce the test set `ISLES` with substantial domain shift from RumourEval datasets. We evaluate top-performing supervised models trained with RumourEval datasets on `ISLES`, also comparing with LLMs in zero-shot ICL. We observe significant performance drop under domain shift. Our ensemble-based framework proposed in Publication I achieves the highest performance among all the supervised methods when evaluated on `ISLES`, although outperformed by LLMs in ICL. We further propose a model-agnostic and data-efficient self-training framework that utilises unlabelled real-world data and synthetic data generated by LLMs for robust model adaptation, achieving better results than

LLMs in zero-shot ICL. We also demonstrate the crucial role of synthetic data quality, and the importance of utilising the unlabelled real-world data to mitigate the differences between synthetic and real-world data.

RQ3: In ICL with LLMs, to what extent does the label design impact the model's ability to generalise to rumour stance classification? How can such generalisation be enhanced through label optimisation? Publication III analyses the impact of three variants in label design (i.e., lexical choice, ordering and elaboration) in zero-shot ICL, revealing that the lexical choice can significantly affect ICL performance on classification tasks. The sensitivity of ICL performance to label ordering is linked to the lexical choice, while elaboration does not show substantial impact. We empirically demonstrate that the ICL performance when varying label names is statistically correlated with the number of outlier neurons in LLM decoder's feed-forward network. We then propose to select optimal label sets that stimulate less outlier neurons in forward pass. Our method demonstrates effectiveness across models, languages, tasks and datasets, including English, French and Portuguese rumour stance classification.

RQ4: How can English-centric models effectively adapt to non-English languages for rumour stance classification? Publication IV investigates adaptation to high- and medium-resource non-English languages. We first establish the SCRum-9 benchmarking dataset, the current largest multilingual rumour stance classification dataset with nine high-resource and medium-resource languages. We then analyse two common approaches for language adaption on SCRum-9: (1) cross-lingual or multilingual fine-tuning MLMs; (2) ICL with LLMs. We observe substantial performance disparities across languages in zero-shot ICL, but normally outperforming MLMs fine-tuned with English or machine-translated multilingual data. We experiment with two strategies to further improve ICL performance, including (1) machine translating target input or demonstration examples; and (2) providing language alignment signal in the prompt. We find that both approaches are effective, with machine translation demonstrating more improvements. We further explore the effectiveness of fine-tuning MLMs with multilingual synthetic data generated by LLMs. It is more computational efficient than ICL, while achieving comparable or even better performance than ICL with the same LLM used for generation.

6.3 Future Work

The conclusions and findings from the above four publications suggest several promising directions for future work.

Multimodal Rumour Stance Classification Rumours in social media platforms are increasingly incorporating images or videos, as they are easier to consume and consequently attract more attention than texts (Alam et al. 2022). Publication I proposes a cross-attention architecture coupled with sample-weighted training objective to enhance the reasoning with textual target rumours. An important direction for future work is to extend this framework to multimodal rumour stance classification, where targets may also contain images or videos. The LLM-assisted self-training framework proposed in Publication II could also be extended

by generating synthetic replies towards text-image rumours, which could support more robust adaptation to real-world scenarios.

Continuous Domain Adaption Publication II introduce a data-efficient method that enables effective adaption to new domains. In real-world application, systems might also be expected to not only deal with new emerging rumours but also long-standing rumours that continue to circulate (Zubiaga et al. 2018*a*). Future work could seek to extend our framework towards continuous unsupervised domain adaption, enabling models to progressively adapt to new rumours and domains while preserving previously acquired knowledge from learnt domains.

Label Set Optimisation for Imbalanced Data Imbalanced data is a known challenge in rumour stance classification as we illustrated in prior work (Li & Scarton 2020) and Publication II. In the LOADS method (Publication III), we assume every stance class contains equal number of samples, which might not hold true in real-world applications. Future work could explore the impact of imbalanced class distribution on LOADS, and further propose label selection methods robust to imbalanced class distributions.

Efficient Multilingual Synthetic Data Selection In Publication IV, we demonstrate the effectiveness of LLM-generated multilingual synthetic data by analysing MLMs' performance after fine-tuning with such data. Our analysis suggests that the effectiveness of synthetic data depends on the choice of MLMs being fine-tuned. Future work could investigate multilingual data quality and characteristics from different analytic perspectives, providing comprehensive insights into optimal LLM selection strategies or data-efficient sampling strategies for more efficiently generating and using multilingual synthetic data.

Improvement for Rumour Collection Notably, one potential limitation of this thesis is the difference between our rumour collection process and that of prior studies (i.e., the RumourEval and PHEME datasets). Specifically, RumourEval and PHEME focus on rumours that are not known a priori. The rumours are identified by journalists who monitor social media platforms in real time. In contrast, the rumours in our `ISLES` and SCRum-9 datasets are sourced from fact-checking websites, which are already verified or debunked at the time of collection. While this design intentionally introduces label shift that features our study in Chapter 3 as well as topical diversity for multilingual analysis in Chapter 4, it also results in a higher proportion of false rumours and deny-stance replies. Consequently, the resulting distribution of rumour veracity and stance expressions may deviate from their natural occurrence in real-world settings. Future work could explore alternative collection strategies that better approximate the real distribution of rumours and stances.

Bibliography

Abu Ahmad, R., Usmanova, A. & Rehm, G. (2025), The ClimateCheck shared task: Scientific fact-checking of social media claims about climate change, *in* T. Ghosal, P. Mayr, A. Singh, A. Naik, G. Rehm, D. Freitag, D. Li, S. Schimmler & A. De Waard, eds, 'Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)', Association for Computational Linguistics, Vienna, Austria, pp. 263–275.
URL: *https://aclanthology.org/2025.sdp-1.24/*

Agerri, R., Centeno, R., Espinosa, M., de Landa, J. F. & Rodrigo, A. (2021), 'Vaxxstanceiberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection', *Procesamiento del lenguaje natural* **67**, 173–181.

Aiyappa, R., Senthilmani, S., An, J., Kwak, H. & Ahn, Y.-Y. (2024), 'Benchmarking zero-shot stance detection with flant5-xxl: Insights from training data, prompting, and decoding strategies into its near-sota performance', *arXiv preprint arXiv:2403.00236* .

Aker, A., Derczynski, L. & Bontcheva, K. (2017), Simple open stance classification for rumour analysis, *in* R. Mitkov & G. Angelova, eds, 'Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017', INCOMA Ltd., Varna, Bulgaria, pp. 31–39.
URL: *https://aclanthology.org/R17-1005/*

Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Martino, G. D. S., Shaar, S., Firooz, H. & Nakov, P. (2022), A survey on multimodal disinformation detection, *in* N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond & S.-H. Na, eds, 'Proceedings of the 29th International Conference on Computational Linguistics', International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp. 6625–6643.
URL: *https://aclanthology.org/2022.coling-1.576/*

Allaway, E. & McKeown, K. (2020), Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations, *in* B. Webber, T. Cohn, Y. He & Y. Liu, eds, 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 8913–8931.
URL: *https://aclanthology.org/2020.emnlp-main.717/*

Aly, R., Guo, Z., Schlichtkrull, M. S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O. & Mittal, A. (2021), The fact extraction and VERification over unstructured

and structured information (FEVEROUS) shared task, *in* R. Aly, C. Christodoulopoulos, O. Cocarascu, Z. Guo, A. Mittal, M. Schlichtkrull, J. Thorne & A. Vlachos, eds, 'Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)', Association for Computational Linguistics, Dominican Republic, pp. 1–13.
URL: *https://aclanthology.org/2021.fever-1.1/*

Antypas, D., Ushio, A., Barbieri, F., Neves, L., Rezaee, K., Espinosa-Anke, L., Pei, J. & Camacho-Collados, J. (2023*a*), SuperTweetEval: A challenging, unified and heterogeneous benchmark for social media NLP research, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Findings of the Association for Computational Linguistics: EMNLP 2023', Association for Computational Linguistics, Singapore, pp. 12590–12607.
URL: *https://aclanthology.org/2023.findings-emnlp.838/*

Antypas, D., Ushio, A., Barbieri, F., Neves, L., Rezaee, K., Espinosa-Anke, L., Pei, J. & Camacho-Collados, J. (2023*b*), 'Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research', *arXiv preprint arXiv:2310.14757* .

Antypas, D., Ushio, A., Camacho-Collados, J., Silva, V., Neves, L. & Barbieri, F. (2022), Twitter topic classification, *in* N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond & S.-H. Na, eds, 'Proceedings of the 29th International Conference on Computational Linguistics', International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp. 3386–3400.
URL: *https://aclanthology.org/2022.coling-1.299/*

Arakelyan, E., Arora, A. & Augenstein, I. (2023), Topic-guided sampling for data-efficient multi-domain stance detection, *in* A. Rogers, J. Boyd-Graber & N. Okazaki, eds, 'Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Toronto, Canada, pp. 13448–13464.
URL: *https://aclanthology.org/2023.acl-long.752/*

Arefyev, N., Sheludko, B., Podolskiy, A. & Panchenko, A. (2020), Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution, *in* D. Scott, N. Bel & C. Zong, eds, 'Proceedings of the 28th International Conference on Computational Linguistics', International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 1242–1255.
URL: *https://aclanthology.org/2020.coling-main.107/*

Atanasova, P., Barron-Cedeno, A., Elsayed, T., Suwaileh, R., Zaghouani, W., Kyuchukov, S., Martino, G. D. S. & Nakov, P. (2018), 'Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness', *arXiv preprint arXiv:1808.05542* .

Augenstein, I., Lioma, C., Wang, D., Chaves Lima, L., Hansen, C., Hansen, C. & Simonsen, J. G. (2019), MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims, *in* K. Inui, J. Jiang, V. Ng & X. Wan, eds, 'Proceedings of the 2019 Conference on

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 4685–4697.
URL: *https://aclanthology.org/D19-1475/*

Bahuleyan, H. & Vechtomova, O. (2017), UWaterloo at SemEval-2017 task 8: Detecting stance towards rumours with topic independent features, *in* S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer & D. Jurgens, eds, 'Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)', Association for Computational Linguistics, Vancouver, Canada, pp. 461–464.
URL: *https://aclanthology.org/S17-2080/*

Bai, F., Ritter, A. & Xu, W. (2021), Pre-train or annotate? domain adaptation with a constrained budget, *in* M.-F. Moens, X. Huang, L. Specia & S. W.-t. Yih, eds, 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 5002–5015.
URL: *https://aclanthology.org/2021.emnlp-main.409/*

Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A. & Slonim, N. (2017), Stance classification of context-dependent claims, *in* M. Lapata, P. Blunsom & A. Koller, eds, 'Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers', Association for Computational Linguistics, Valencia, Spain, pp. 251–261.
URL: *https://aclanthology.org/E17-1024/*

Barbieri, F., Espinosa Anke, L. & Camacho-Collados, J. (2022), XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond, *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk & S. Piperidis, eds, 'Proceedings of the Thirteenth Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 258–266.
URL: *https://aclanthology.org/2022.lrec-1.27/*

Barriere, V., Jacquet, G. G. & Hemamou, L. (2022), Cofe: A new dataset of intra-multilingual multi-target stance classification from an online european participatory democracy platform, *in* 'Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)', pp. 418–422.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**, 5–32.

Chen, D., Wang, D., Darrell, T. & Ebrahimi, S. (2022), Contrastive test-time adaptation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 295–305.

Chen, S., Khashabi, D., Yin, W., Callison-Burch, C. & Roth, D. (2019), Seeing things from a different angle:discovering diverse perspectives about claims, *in* J. Burstein, C. Doran &

T. Solorio, eds, 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 542–557.
URL: *https://aclanthology.org/N19-1053/*

Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X. & Wang, W. Y. (2020), Tabfact: A large-scale dataset for table-based fact verification, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=rkeJRhNYDH*

Chen, Y., Yuan, L., Cui, G., Liu, Z. & Ji, H. (2023), A close look into the calibration of pre-trained language models, *in* A. Rogers, J. Boyd-Graber & N. Okazaki, eds, 'Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Toronto, Canada, pp. 1343–1367.
URL: *https://aclanthology.org/2023.acl-long.75/*

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S. et al. (2024), 'Scaling instruction-finetuned language models', *Journal of Machine Learning Research* **25**(70), 1–53.

Clark, C., Yatskar, M. & Zettlemoyer, L. (2019), Don't take the easy way out: Ensemble based methods for avoiding known dataset biases, *in* K. Inui, J. Jiang, V. Ng & X. Wan, eds, 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 4069–4082.
URL: *https://aclanthology.org/D19-1418/*

Clark, T., Conforti, C., Liu, F., Meng, Z., Shareghi, E. & Collier, N. (2021), Integrating transformers and knowledge graphs for Twitter stance detection, *in* W. Xu, A. Ritter, T. Baldwin & A. Rahimi, eds, 'Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)', Association for Computational Linguistics, Online, pp. 304–312.
URL: *https://aclanthology.org/2021.wnut-1.34/*

Conforti, C., Berndt, J., Pilehvar, M. T., Giannitsarou, C., Toxvaerd, F. & Collier, N. (2020), Will-they-won't-they: A very large dataset for stance detection on Twitter, *in* D. Jurafsky, J. Chai, N. Schluter & J. Tetreault, eds, 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 1715–1724.
URL: *https://aclanthology.org/2020.acl-main.157/*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020), Unsupervised cross-lingual representation learning at scale, *in* D. Jurafsky, J. Chai, N. Schluter & J. Tetreault, eds, 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 8440–8451.
URL: *https://aclanthology.org/2020.acl-main.747/*

Cook, O., Grimshaw, C., Wu, B. P., Dillon, S., Hicks, J., Jones, L., Smith, T., Szert, M. & Song, X. (2025), Efficient annotator reliability assessment and sample weighting for knowledge-based misinformation detection on social media, *in* L. Chiruzzo, A. Ritter & L. Wang, eds, 'Findings of the Association for Computational Linguistics: NAACL 2025', Association for Computational Linguistics, Albuquerque, New Mexico, pp. 3348–3358.
URL: *https://aclanthology.org/2025.findings-naacl.185/*

Cui, G., Hu, S., Ding, N., Huang, L. & Liu, Z. (2022), Prototypical verbalizer for prompt-based few-shot tuning, *in* S. Muresan, P. Nakov & A. Villavicencio, eds, 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Dublin, Ireland, pp. 7014–7024.
URL: *https://aclanthology.org/2022.acl-long.483/*

Cui, S., Wang, S., Zhuo, J., Li, L., Huang, Q. & Tian, Q. (2020), Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 3941–3950.

Daumé III, H. (2007), Frustratingly easy domain adaptation, *in* A. Zaenen & A. van den Bosch, eds, 'Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics', Association for Computational Linguistics, Prague, Czech Republic, pp. 256–263.
URL: *https://aclanthology.org/P07-1033/*

Dawid, A. P. & Skene, A. M. (1979), 'Maximum Likelihood Estimation of Observer Error-rates using the EM Algorithm', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **28**(1), 20–28.
URL: *https://www.jstor.org/stable/2346806*

Deng, M., Wang, J., Hsieh, C.-P., Wang, Y., Guo, H., Shu, T., Song, M., Xing, E. & Hu, Z. (2022), RLPrompt: Optimizing discrete text prompts with reinforcement learning, *in* Y. Goldberg, Z. Kozareva & Y. Zhang, eds, 'Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 3369–3391.
URL: *https://aclanthology.org/2022.emnlp-main.222/*

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G. & Zubiaga, A. (2017), SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours, *in* S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer & D. Jurgens, eds, 'Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)', Association for Computational Linguistics, Vancouver, Canada, pp. 69–76.
URL: *https://aclanthology.org/S17-2006/*

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *in* J. Burstein, C. Doran & T. Solorio, eds, 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short

Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.
URL: *https://aclanthology.org/N19-1423/*

Ding, N., Xu, Y., Tang, Y., Xu, C., Wang, Y. & Tao, D. (2022), Source-free domain adaptation via distribution estimation, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 7212–7222.

Dougrez-Lewis, J., Liakata, M., Kochkina, E. & He, Y. (2021), Learning disentangled latent topics for twitter rumour veracity classification, *in* 'Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021', pp. 3902–3908.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A. et al. (2024), 'The llama 3 herd of models', *arXiv preprint arXiv:2407.21783* .

Dumitrache, A., Inel, O., Aroyo, L., Timmermans, B. & Welty, C. (2018), 'Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement', *arXiv preprint arXiv:1808.06080* .

Dungs, S., Aker, A., Fuhr, N. & Bontcheva, K. (2018), Can rumour stance alone predict veracity?, *in* E. M. Bender, L. Derczynski & P. Isabelle, eds, 'Proceedings of the 27th International Conference on Computational Linguistics', Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 3360–3370.
URL: *https://aclanthology.org/C18-1284/*

Elfwing, S., Uchibe, E. & Doya, K. (2018), 'Sigmoid-weighted linear units for neural network function approximation in reinforcement learning', *Neural networks* **107**, 3–11.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S. & Olah, C. (2021), 'A mathematical framework for transformer circuits', *Transformer Circuits Thread* . https://transformer-circuits.pub/2021/framework/index.html.

España-Bonet, C. (2023), Multilingual coarse political stance classification of media. the editorial line of a ChatGPT and bard newspaper, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Findings of the Association for Computational Linguistics: EMNLP 2023', Association for Computational Linguistics, Singapore, pp. 11757–11777.
URL: *https://aclanthology.org/2023.findings-emnlp.787/*

Fajcik, M., Smrz, P. & Burget, L. (2019), BUT-FIT at SemEval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers, *in* J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki & S. M. Mohammad, eds, 'Proceedings of the 13th International Workshop on Semantic Evaluation', Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 1097–1104.
URL: *https://aclanthology.org/S19-2192/*

Fang, Y., Yap, P.-T., Lin, W., Zhu, H. & Liu, M. (2024), 'Source-free unsupervised domain adaptation: A survey', *Neural Networks* p. 106230.

Fellbaum, C. (1998), 'Wordnet: An electronic lexical database', *MIT Press google schola* **2**, 678–686.

Ferreira, W. & Vlachos, A. (2016), Emergent: a novel data-set for stance classification, *in* K. Knight, A. Nenkova & O. Rambow, eds, 'Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, San Diego, California, pp. 1163–1168.
URL: *https://aclanthology.org/N16-1138/*

Ganin, Y. & Lempitsky, V. (2015), Unsupervised domain adaptation by backpropagation, *in* 'International conference on machine learning', PMLR, pp. 1180–1189.

Gao, T., Fisch, A. & Chen, D. (2021), Making pre-trained language models better few-shot learners, *in* C. Zong, F. Xia, W. Li & R. Navigli, eds, 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', Association for Computational Linguistics, Online, pp. 3816–3830.
URL: *https://aclanthology.org/2021.acl-long.295/*

García Lozano, M., Lilja, H., Tjörnhammar, E. & Karasalo, M. (2017), Mama edha at SemEval-2017 task 8: Stance classification with CNN and rules, *in* S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer & D. Jurgens, eds, 'Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)', Association for Computational Linguistics, Vancouver, Canada, pp. 481–485.
URL: *https://aclanthology.org/S17-2084/*

Gatto, J., Sharif, O. & Preum, S. (2023), Chain-of-thought embeddings for stance detection on social media, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Findings of the Association for Computational Linguistics: EMNLP 2023', Association for Computational Linguistics, Singapore, pp. 4154–4161.
URL: *https://aclanthology.org/2023.findings-emnlp.273/*

Gencheva, P., Nakov, P., Màrquez, L., Barrón-Cedeño, A. & Koychev, I. (2017), A context-aware approach for detecting worth-checking claims in political debates, *in* R. Mitkov & G. Angelova, eds, 'Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017', INCOMA Ltd., Varna, Bulgaria, pp. 267–276.
URL: *https://aclanthology.org/R17-1037/*

Geva, M., Caciularu, A., Wang, K. & Goldberg, Y. (2022), Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space, *in* Y. Goldberg, Z. Kozareva & Y. Zhang, eds, 'Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 30–45.
URL: *https://aclanthology.org/2022.emnlp-main.3/*

Ghanem, B., Cignarella, A. T., Bosco, C., Rosso, P. & Rangel Pardo, F. M. (2019), UPV-28-UNITO at SemEval-2019 task 7: Exploiting post's nesting and syntax information for rumor

stance classification, *in* J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki & S. M. Mohammad, eds, 'Proceedings of the 13th International Workshop on Semantic Evaluation', Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 1125–1131.
URL: *https://aclanthology.org/S19-2197/*

Glavaš, G., Šnajder, J., Kordjamshidi, P., Moens, M.-F., Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J. et al. (2014), Hieve: A corpus for extracting event hierarchies from news stories, *in* 'Proceedings of 9th language resources and evaluation conference', ELRA, pp. 3678–3683.

Gonen, H., Iyer, S., Blevins, T., Smith, N. & Zettlemoyer, L. (2023), Demystifying prompts in language models via perplexity estimation, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Findings of the Association for Computational Linguistics: EMNLP 2023', Association for Computational Linguistics, Singapore, pp. 10136–10148.
URL: *https://aclanthology.org/2023.findings-emnlp.679/*

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K. & Derczynski, L. (2019*a*), SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours, *in* J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki & S. M. Mohammad, eds, 'Proceedings of the 13th International Workshop on Semantic Evaluation', Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 845–854.
URL: *https://aclanthology.org/S19-2147/*

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K. & Derczynski, L. (2019*b*), Semeval-2019 task 7: Rumoureval, determining rumour veracity and support for rumours, *in* 'Proceedings of the 13th International Workshop on Semantic Evaluation', pp. 845–854.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A. et al. (2024), 'The llama 3 herd of models', *arXiv preprint arXiv:2407.21783* .

Grootendorst, M. (2022), 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure'.
URL: *https://arxiv.org/abs/2203.05794*

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. et al. (2025), 'Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning', *arXiv preprint arXiv:2501.12948* .

Guo, Z., Schlichtkrull, M. & Vlachos, A. (2022), 'A survey on automated fact-checking', *Transactions of the Association for Computational Linguistics* **10**, 178–206.
URL: *https://aclanthology.org/2022.tacl-1.11/*

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D. & Smith, N. A. (2020), Don't stop pretraining: Adapt language models to domains and tasks, *in* D. Jurafsky, J. Chai, N. Schluter & J. Tetreault, eds, 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 8342–8360.
URL: *https://aclanthology.org/2020.acl-main.740/*

Hambardzumyan, K., Khachatrian, H. & May, J. (2021), WARP: Word-level Adversarial ReProgramming, *in* C. Zong, F. Xia, W. Li & R. Navigli, eds, 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', Association for Computational Linguistics, Online, pp. 4921–4933.
URL: *https://aclanthology.org/2021.acl-long.381/*

Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T. T. H., Hackl, G., Moreno, J. G. & Doucet, A. (2021), A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers, *in* 'Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval', pp. 2328–2334.

Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M. & Gurevych, I. (2018), A retrospective analysis of the fake news challenge stance-detection task, *in* E. M. Bender, L. Derczynski & P. Isabelle, eds, 'Proceedings of the 27th International Conference on Computational Linguistics', Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 1859–1874.
URL: *https://aclanthology.org/C18-1158/*

Hanselowski, A., Stab, C., Schulz, C., Li, Z. & Gurevych, I. (2019), A richly annotated corpus for different tasks in automated fact-checking, *in* M. Bansal & A. Villavicencio, eds, 'Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)', Association for Computational Linguistics, Hong Kong, China, pp. 493–503.
URL: *https://aclanthology.org/K19-1046/*

Haouari, F., Scarton, C., Faggiani, N., Nikolaidis, N., Kotseva, B., Farha, I. A., Linge, J. & Bontcheva, K. (2025), Ukelectionnarratives: A dataset of misleading narratives surrounding recent uk general elections, *in* 'Proceedings of the International AAAI Conference on Web and Social Media', Vol. 19, pp. 2477–2495.

Hardalov, M., Arora, A., Nakov, P. & Augenstein, I. (2021), Cross-domain label-adaptive stance detection, *in* M.-F. Moens, X. Huang, L. Specia & S. W.-t. Yih, eds, 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 9011–9028.
URL: *https://aclanthology.org/2021.emnlp-main.710/*

Hardalov, M., Arora, A., Nakov, P. & Augenstein, I. (2022*a*), Few-shot cross-lingual stance detection with sentiment-based pre-training, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 36, pp. 10729–10737.

Hardalov, M., Arora, A., Nakov, P. & Augenstein, I. (2022*b*), A survey on stance detection for mis- and disinformation identification, *in* M. Carpuat, M.-C. de Marneffe & I. V. Meza Ruiz, eds, 'Findings of the Association for Computational Linguistics: NAACL 2022', Association for Computational Linguistics, Seattle, United States, pp. 1259–1277.
URL: *https://aclanthology.org/2022.findings-naacl.94/*

Hardalov, M., Arora, A., Nakov, P. & Augenstein, I. (2022*c*), A survey on stance detection for mis-and disinformation identification, *in* 'Findings of the Association for Computational

Linguistics: NAACL 2022', pp. 1259–1277.
URL: *https://aclanthology.org/2022.findings-naacl.94/*

Hasan, K. S. & Ng, V. (2013), Stance classification of ideological debates: Data, models, features, and constraints, *in* R. Mitkov & J. C. Park, eds, 'Proceedings of the Sixth International Joint Conference on Natural Language Processing', Asian Federation of Natural Language Processing, Nagoya, Japan, pp. 1348–1356.
URL: *https://aclanthology.org/I13-1191/*

He, Z., Mokhberian, N. & Lerman, K. (2022), Infusing knowledge from Wikipedia to enhance stance detection, *in* J. Barnes, O. De Clercq, V. Barriere, S. Tafreshi, S. Alqahtani, J. Sedoc, R. Klinger & A. Balahur, eds, 'Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis', Association for Computational Linguistics, Dublin, Ireland, pp. 71–77.
URL: *https://aclanthology.org/2022.wassa-1.7/*

Hendrycks, D. & Gimpel, K. (2016), 'Gaussian error linear units (gelus)', *arXiv preprint arXiv:1606.08415* .

Hong, J., Zhang, Y.-D. & Chen, W. (2022), 'Source-free unsupervised domain adaptation for cross-modality abdominal multi-organ segmentation', *Knowledge-Based Systems* **250**, 109155.

Hossain, T., Logan IV, R. L., Ugarte, A., Matsubara, Y., Young, S. & Singh, S. (2020), COVIDLies: Detecting COVID-19 misinformation on social media, *in* K. Verspoor, K. B. Cohen, M. Conway, B. de Bruijn, M. Dredze, R. Mihalcea & B. Wallace, eds, 'Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020', Association for Computational Linguistics, Online.
URL: *https://aclanthology.org/2020.nlpcovid19-2.11/*

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L. et al. (2023), 'Mistral 7b', *arXiv preprint arXiv:2310.06825* .

Jiang, Y., Song, X., Scarton, C., Singh, I., Aker, A. & Bontcheva, K. (2023), Categorising fine-to-coarse grained misinformation: An empirical study of the COVID-19 infodemic, *in* R. Mitkov & G. Angelova, eds, 'Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing', INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, pp. 556–567.
URL: *https://aclanthology.org/2023.ranlp-1.61/*

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. & Amodei, D. (2020), 'Scaling laws for neural language models', *arXiv preprint arXiv:2001.08361* .

Kaushal, A., Saha, A. & Ganguly, N. (2021), tWT–WT: A dataset to assert the role of target entities for detecting stance of tweets, *in* K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou, eds, 'Proceedings of the 2021 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Online, pp. 3879–3889.
URL: *https://aclanthology.org/2021.naacl-main.303/*

Kim, H. J., Cho, H., Kim, J., Kim, T., Yoo, K. M. & Lee, S.-g. (2022), 'Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator', *arXiv preprint arXiv:2206.08082* .

Kim, Y., Cho, D., Han, K., Panda, P. & Hong, S. (2021), 'Domain adaptation without source data', *IEEE Transactions on Artificial Intelligence* **2**(6), 508–518.

Klingner, M., Termöhlen, J.-A., Ritterbach, J. & Fingscheidt, T. (2022), Unsupervised batchnorm adaptation (ubna): A domain adaptation method for semantic segmentation without using source domain representations, *in* 'Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision', pp. 210–220.

Kobbe, J., Hulpuș, I. & Stuckenschmidt, H. (2020), Unsupervised stance detection for arguments from consequences, *in* B. Webber, T. Cohn, Y. He & Y. Liu, eds, 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 50–60.
URL: *https://aclanthology.org/2020.emnlp-main.4/*

Kochkina, E., Hossain, T., Logan IV, R. L., Arana-Catania, M., Procter, R., Zubiaga, A., Singh, S., He, Y. & Liakata, M. (2023), 'Evaluating the generalisability of neural rumour verification models', *Information Processing & Management* **60**(1), 103116.

Kochkina, E. & Liakata, M. (2020), Estimating predictive uncertainty for rumour verification models, *in* D. Jurafsky, J. Chai, N. Schluter & J. Tetreault, eds, 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 6964–6981.
URL: *https://aclanthology.org/2020.acl-main.623/*

Kochkina, E., Liakata, M. & Augenstein, I. (2017), Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM, *in* S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer & D. Jurgens, eds, 'Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)', Association for Computational Linguistics, Vancouver, Canada, pp. 475–480.
URL: *https://aclanthology.org/S17-2083/*

Kochkina, E., Liakata, M. & Zubiaga, A. (2018), All-in-one: Multi-task learning for rumour verification, *in* E. M. Bender, L. Derczynski & P. Isabelle, eds, 'Proceedings of the 27th International Conference on Computational Linguistics', Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 3402–3413.
URL: *https://aclanthology.org/C18-1288/*

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R. et al. (2023), 'Openassistant conversations–democratizing large language model alignment', *arXiv preprint arXiv:2304.07327* .

Küçük, D. & Can, F. (2020), 'Stance detection: A survey', *ACM Computing Surveys (CSUR)* **53**(1), 1–37.

Kumar, H., Shah, J., Hegde, N., Gupta, P., Jindal, V. & Modi, A. (2021), IITK at SemEval-2021 task 10: Source-free unsupervised domain adaptation using class prototypes, *in* A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot & X. Zhu, eds, 'Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)', Association for Computational Linguistics, Online, pp. 438–444.
URL: *https://aclanthology.org/2021.semeval-1.53/*

Kumar, S. & Carley, K. (2019), Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations, *in* A. Korhonen, D. Traum & L. Màrquez, eds, 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 5047–5058.
URL: *https://aclanthology.org/P19-1498/*

Kurmi, V. K., Subramanian, V. K. & Namboodiri, V. P. (2021), Domain impression: A source data free domain adaptation method, *in* 'Proceedings of the IEEE/CVF winter conference on applications of computer vision', pp. 615–625.

Kurniawan, K., Frermann, L., Schulz, P. & Cohn, T. (2021), PTST-UoM at SemEval-2021 task 10: Parsimonious transfer for sequence tagging, *in* A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot & X. Zhu, eds, 'Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)', Association for Computational Linguistics, Online, pp. 445–451.
URL: *https://aclanthology.org/2021.semeval-1.54/*

Kurz, S., Chen, J.-J., Flek, L. & Zhao, Z. (2024), 'Investigating language-specific calibration for pruning multilingual large language models', *arXiv preprint arXiv:2408.14398* .

Kuzmin, A., Nagel, M., Baalen, M. V., Behboodi, A. & Blankevoort, T. (2023), Pruning vs quantization: Which is better?, *in* 'Thirty-seventh Conference on Neural Information Processing Systems'.
URL: *https://openreview.net/forum?id=0OU1ZXXxs5*

Lan, X., Gao, C., Jin, D. & Li, Y. (2024), 'Stance detection with collaborative role-infused llm-based agents', *Proceedings of the International AAAI Conference on Web and Social Media* **18**(1), 891–903.
URL: *https://ojs.aaai.org/index.php/ICWSM/article/view/31360*

Laparra, E., Su, X., Zhao, Y., Uzuner, Ö., Miller, T. & Bethard, S. (2021), SemEval-2021 task 10: Source-free domain adaptation for semantic processing, *in* A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot & X. Zhu, eds, 'Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)', Association for Computational Linguistics, Online, pp. 348–356.
URL: *https://aclanthology.org/2021.semeval-1.42/*

Lavrouk, A., Ligon, I., Zheng, J., Naous, T., Xu, W. & Ritter, A. (2024), Stanceosaurus 2.0 - classifying stance towards Russian and Spanish misinformation, *in* R. van der Goot, J. Bak,

M. Müller-Eberstein, W. Xu, A. Ritter & T. Baldwin, eds, 'Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)', Association for Computational Linguistics, San Ġiljan, Malta, pp. 31–43.
URL: *https://aclanthology.org/2024.wnut-1.4/*

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M. et al. (2009), 'Life in the network: the coming age of computational social science', *Science (New York, NY)* **323**(5915), 721.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2745217/*

Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M. et al. (2023), 'Bloom: A 176b-parameter open-access multilingual language model'.

Le Scao, T. & Rush, A. (2021), How many data points is a prompt worth?, *in* K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou, eds, 'Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Online, pp. 2627–2636.
URL: *https://aclanthology.org/2021.naacl-main.208/*

Lester, B., Al-Rfou, R. & Constant, N. (2021), The power of scale for parameter-efficient prompt tuning, *in* M.-F. Moens, X. Huang, L. Specia & S. W.-t. Yih, eds, 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 3045–3059.
URL: *https://aclanthology.org/2021.emnlp-main.243/*

Li, J., Chen, E., Ding, Z., Zhu, L., Lu, K. & Shen, H. T. (2020), 'Maximum density divergence for domain adaptation', *IEEE transactions on pattern analysis and machine intelligence* **43**(11), 3918–3930.

Li, J., Sujana, Y. & Kao, H.-Y. (2020), Exploiting microblog conversation structures to detect rumors, *in* D. Scott, N. Bel & C. Zong, eds, 'Proceedings of the 28th International Conference on Computational Linguistics', International Committee on Computational Linguistics, Barcelona, Spain (Online), pp. 5420–5429.
URL: *https://aclanthology.org/2020.coling-main.473/*

Li, Q., Zhang, Q. & Si, L. (2019), eventAI at SemEval-2019 task 7: Rumor detection on social media by exploiting content, user credibility and propagation information, *in* J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki & S. M. Mohammad, eds, 'Proceedings of the 13th International Workshop on Semantic Evaluation', Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 855–859.
URL: *https://aclanthology.org/S19-2148/*

Li, R., Jiao, Q., Cao, W., Wong, H.-S. & Wu, S. (2020), Model adaptation: Unsupervised domain adaptation without source data, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 9641–9650.

Li, Y. & Scarton, C. (2020), Revisiting rumour stance classification: Dealing with imbalanced data, *in* A. Aker & A. Zubiaga, eds, 'Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)', Association for Computational Linguistics, Barcelona, Spain (Online), pp. 38–44.
URL: *https://aclanthology.org/2020.rdsm-1.4/*

Li, Y. & Scarton, C. (2024), Can we identify stance without target arguments? a study for rumour stance classification, *in* N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue, eds, 'Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)', ELRA and ICCL, Torino, Italia, pp. 2844–2851.
URL: *https://aclanthology.org/2024.lrec-main.253/*

Li, Y., Scarton, C., Song, X. & Bontcheva, K. (2023), Classifying COVID-19 vaccine narratives, *in* R. Mitkov & G. Angelova, eds, 'Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing', INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, pp. 648–657.
URL: *https://aclanthology.org/2023.ranlp-1.70/*

Li, Y., Vasilakes, J., Zhao, Z. & Scarton, C. (2025), 'Scrum-9: Multilingual stance classification over rumours on social media', *arXiv preprint arXiv:2505.18916* .

Li, Y., Zhao, C. & Caragea, C. (2021), Improving stance detection with multi-dataset learning and knowledge distillation, *in* M.-F. Moens, X. Huang, L. Specia & S. W.-t. Yih, eds, 'Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 6332–6345.
URL: *https://aclanthology.org/2021.emnlp-main.511/*

Li, Y., Zhao, Z. & Scarton, C. (2024), 'Label set optimization via activation distribution kurtosis for zero-shot classification with generative models', *arXiv preprint arXiv:2410.19195* .

Li, Y., Zhao, Z. & Scarton, C. (2025), 'It's all about in-context learning! teaching extremely low-resource languages to llms', *arXiv preprint arXiv:2508.19089* .

Liang, B., Chen, Z., Gui, L., He, Y., Yang, M. & Xu, R. (2022), Zero-shot stance detection via contrastive learning, *in* 'Proceedings of the ACM Web Conference 2022', pp. 2738–2747.

Liang, J., Hu, D. & Feng, J. (2020), Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation, *in* 'International conference on machine learning', PMLR, pp. 6028–6039.

Lillie, A. E., Middelboe, E. R. & Derczynski, L. (2019*a*), Joint rumour stance and veracity prediction, *in* M. Hartmann & B. Plank, eds, 'Proceedings of the 22nd Nordic Conference on Computational Linguistics', Linköping University Electronic Press, Turku, Finland, pp. 208–221.
URL: *https://aclanthology.org/W19-6122/*

Lillie, A. E., Middelboe, E. R. & Derczynski, L. (2019*b*), Joint rumour stance and veracity prediction, *in* 'Nordic Conference of Computational Linguistics (2019)', Linköping University Electronic Press, pp. 208–221.

Lin, Z. & Lee, K. (2024), Dual operating modes of in-context learning, *in* 'Forty-first International Conference on Machine Learning'.
URL: *https://openreview.net/forum?id=ElVHUWyL3n*

Liu, H., Das, A., Boltz, A., Zhou, D., Pinaroc, D., Lease, M. & Lee, M. K. (2024), 'Human-centered nlp fact-checking: Co-designing with fact-checkers using matchmaking for ai', *Proceedings of the ACM on Human-Computer Interaction* **8**(CSCW2), 1–44.

Liu, H., Tam, D., Mohammed, M., Mohta, J., Huang, T., Bansal, M. & Raffel, C. (2022), Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, *in* A. H. Oh, A. Agarwal, D. Belgrave & K. Cho, eds, 'Advances in Neural Information Processing Systems'.
URL: *https://openreview.net/forum?id=rBCvMG-JsPd*

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L. & Chen, W. (2022), What makes good in-context examples for GPT-3?, *in* E. Agirre, M. Apidianaki & I. Vulić, eds, 'Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures', Association for Computational Linguistics, Dublin, Ireland and Online, pp. 100–114.
URL: *https://aclanthology.org/2022.deelio-1.10/*

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. & Liang, P. (2024), 'Lost in the middle: How language models use long contexts', *Transactions of the Association for Computational Linguistics* **12**, 157–173.
URL: *https://aclanthology.org/2024.tacl-1.9/*

Liu, R., Lin, Z., Ji, H., Li, J., Fu, P. & Wang, W. (2022), Target really matters: Target-aware contrastive learning and consistency regularization for few-shot stance detection, *in* N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond & S.-H. Na, eds, 'Proceedings of the 29th International Conference on Computational Linguistics', International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp. 6944–6954.
URL: *https://aclanthology.org/2022.coling-1.605/*

Liu, S., Xing, L. & Zou, J. (2023), 'In-context vectors: Making in context learning more effective and controllable through latent space steering', *arXiv preprint arXiv:2311.06668* .

Liu, X., Ji, K., Fu, Y., Tam, W., Du, Z., Yang, Z. & Tang, J. (2022), P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks, *in* S. Muresan, P. Nakov & A. Villavicencio, eds, 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)', Association for Computational Linguistics, Dublin, Ireland, pp. 61–68.
URL: *https://aclanthology.org/2022.acl-short.8/*

Liu, X., Xing, F., Yang, C., El Fakhri, G. & Woo, J. (2021), Adapting off-the-shelf source segmenter for target medical image segmentation, *in* 'Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24', Springer, pp. 549–559.

Liu, Y., Liu, J., Shi, X., Cheng, Q. & Lu, W. (2024), 'Let's learn step by step: Enhancing in-context learning ability with curriculum learning', *arXiv preprint arXiv:2402.10738* .

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019), 'Roberta: A robustly optimized bert pretraining approach', *arXiv preprint arXiv:1907.11692* .
URL: *https://arxiv.org/abs/1907.11692*

Long, M., Cao, Y., Wang, J. & Jordan, M. (2015), Learning transferable features with deep adaptation networks, *in* 'International conference on machine learning', PMLR, pp. 97–105.

Long, M., Cao, Z., Wang, J. & Jordan, M. I. (2018), 'Conditional adversarial domain adaptation', *Advances in neural information processing systems* **31**.

Loshchilov, I. & Hutter, F. (2017), 'Decoupled weight decay regularization', *arXiv preprint arXiv:1711.05101* .

Loshchilov, I. & Hutter, F. (2019), Decoupled weight decay regularization, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=Bkg6RiCqY7*

Lozhnikov, N., Derczynski, L. & Mazzara, M. (2018), Stance prediction for russian: data and analysis, *in* 'International Conference in Software Engineering for Defence Applications', Springer, pp. 176–186.

Lu, Y., Bartolo, M., Moore, A., Riedel, S. & Stenetorp, P. (2022), Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, *in* S. Muresan, P. Nakov & A. Villavicencio, eds, 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Dublin, Ireland, pp. 8086–8098.
URL: *https://aclanthology.org/2022.acl-long.556/*

Malik, B., Ramesh Kashyap, A., Kan, M.-Y. & Poria, S. (2023), UDAPTER - efficient domain adaptation using adapters, *in* A. Vlachos & I. Augenstein, eds, 'Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics', Association for Computational Linguistics, Dubrovnik, Croatia, pp. 2249–2263.
URL: *https://aclanthology.org/2023.eacl-main.165/*

Mao, J., Middleton, S. E. & Niranjan, M. (2024), Do prompt positions really matter?, *in* K. Duh, H. Gomez & S. Bethard, eds, 'Findings of the Association for Computational Linguistics: NAACL 2024', Association for Computational Linguistics, Mexico City, Mexico, pp. 4102–4130.
URL: *https://aclanthology.org/2024.findings-naacl.258/*

Miao, Z., Li, Y., Wang, X. & Tan, W.-C. (2020), Snippext: Semi-supervised opinion mining with augmented data, *in* 'Proceedings of The Web Conference 2020', pp. 617–628.

Mihaylova, T., Nakov, P., Màrquez, L., Barrón-Cedeño, A., Mohtarami, M., Karadzhov, G. & Glass, J. (2018), Fact checking in community forums, *in* 'Proceedings of the AAAI conference on artificial intelligence', Vol. 32.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013), 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I. & Roth, D. (2023), 'Recent advances in natural language processing via large pre-trained language models: A survey', *ACM Computing Surveys* **56**(2), 1–40.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X. & Cherry, C. (2016), SemEval-2016 task 6: Detecting stance in tweets, *in* S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov & T. Zesch, eds, 'Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)', Association for Computational Linguistics, San Diego, California, pp. 31–41.
URL: *https://aclanthology.org/S16-1003/*

Mu, Y., Jin, M., Grimshaw, C., Scarton, C., Bontcheva, K. & Song, X. (2023), Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter, *in* 'Proceedings of the International AAAI Conference on Web and Social Media', Vol. 17, pp. 1052–1062.

Mu, Y., Wu, B. P., Thorne, W., Robinson, A., Aletras, N., Scarton, C., Bontcheva, K. & Song, X. (2024), Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science, *in* N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue, eds, 'Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)', ELRA and ICCL, Torino, Italia, pp. 12074–12086.
URL: *https://aclanthology.org/2024.lrec-main.1055/*

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., Papotti, P., Shaar, S. & Da San Martino, G. (2021), Automated fact-checking for assisting human fact-checkers, *in* Z.-H. Zhou, ed., 'Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21', International Joint Conferences on Artificial Intelligence Organization, pp. 4551–4558. Survey Track.

Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeno, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N. et al. (2021), The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, *in* 'European Conference on Information Retrieval', Springer, pp. 639–649.

Nguyen, D. Q., Vu, T. & Tuan Nguyen, A. (2020), BERTweet: A pre-trained language model for English tweets, *in* Q. Liu & D. Schlangen, eds, 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', Association for Computational Linguistics, Online, pp. 9–14.
URL: *https://aclanthology.org/2020.emnlp-demos.2/*

Nikolaidis, N., Stefanovitch, N., Silvano, P., Dimitrov, D. I., Yangarber, R., Guimarães, N., Sartori, E., Androutsopoulos, I., Nakov, P., Da San Martino, G. & Piskorski, J. (2025), PolyNarrative: A multilingual, multilabel, multi-domain dataset for narrative extraction from news articles, *in* W. Che, J. Nabende, E. Shutova & M. T. Pilehvar, eds, 'Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Vienna, Austria, pp. 31323–31345.
URL: *https://aclanthology.org/2025.acl-long.1513/*

Niven, T. & Kao, H.-Y. (2019), Probing neural network comprehension of natural language arguments, *in* A. Korhonen, D. Traum & L. Màrquez, eds, 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 4658–4664.
URL: *https://aclanthology.org/P19-1459/*

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022), 'Training language models to follow instructions with human feedback', *Advances in neural information processing systems* **35**, 27730–27744.

Peinelt, N., Liakata, M. & Nguyen, D. (2019), Aiming beyond the obvious: Identifying non-obvious cases in semantic similarity datasets, *in* A. Korhonen, D. Traum & L. Màrquez, eds, 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 2792–2798.
URL: *https://aclanthology.org/P19-1268/*

Peng, K., Ding, L., Yuan, Y., Liu, X., Zhang, M., Ouyang, Y. & Tao, D. (2024), Revisiting demonstration selection strategies in in-context learning, *in* L.-W. Ku, A. Martins & V. Srikumar, eds, 'Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Bangkok, Thailand, pp. 9090–9101.
URL: *https://aclanthology.org/2024.acl-long.492/*

Pennington, J., Socher, R. & Manning, C. D. (2014), Glove: Global vectors for word representation, *in* 'Empirical Methods in Natural Language Processing (EMNLP)', pp. 1532–1543.
URL: *http://www.aclweb.org/anthology/D14-1162*

Phuvipadawat, S. & Murata, T. (2010), Breaking news detection and tracking in twitter, *in* '2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology', Vol. 3, IEEE, pp. 120–123.

Pires, T., Schlinger, E. & Garrette, D. (2019), How multilingual is multilingual BERT?, *in* A. Korhonen, D. Traum & L. Màrquez, eds, 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 4996–5001.
URL: *https://aclanthology.org/P19-1493/*

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R. & Van Durme, B. (2018), Hypothesis only baselines in natural language inference, *in* M. Nissim, J. Berant & A. Lenci, eds,

'Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics', Association for Computational Linguistics, New Orleans, Louisiana, pp. 180–191.
URL: *https://aclanthology.org/S18-2023/*

Qazvinian, V., Rosengren, E., Radev, D. R. & Mei, Q. (2011), Rumor has it: Identifying misinformation in microblogs, *in* R. Barzilay & M. Johnson, eds, 'Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Edinburgh, Scotland, UK., pp. 1589–1599.
URL: *https://aclanthology.org/D11-1147/*

Razumovskaia, E., Vulić, I. & Korhonen, A. (2025), 'Analyzing and adapting large language models for few-shot multilingual nlu: Are we there yet?', *Transactions of the Association for Computational Linguistics* **13**, 1096–1120.
URL: *https://doi.org/10.1162/TACL.a.33*

Reimers, N. & Gurevych, I. (2019), Sentence-BERT: Sentence embeddings using Siamese BERT-networks, *in* K. Inui, J. Jiang, V. Ng & X. Wan, eds, 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 3982–3992.
URL: *https://aclanthology.org/D19-1410/*

Rubin, O., Herzig, J. & Berant, J. (2022), Learning to retrieve prompts for in-context learning, *in* M. Carpuat, M.-C. de Marneffe & I. V. Meza Ruiz, eds, 'Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Seattle, United States, pp. 2655–2671.
URL: *https://aclanthology.org/2022.naacl-main.191/*

Saakyan, A., Chakrabarty, T. & Muresan, S. (2021), COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic, *in* C. Zong, F. Xia, W. Li & R. Navigli, eds, 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', Association for Computational Linguistics, Online, pp. 2116–2129.
URL: *https://aclanthology.org/2021.acl-long.165/*

Scarton, C. & Li, Y. (2021), Cross-lingual rumour stance classification: a first study with bert and machine translation., *in* 'TTO', pp. 50–59.

Scarton, C., Silva, D. & Bontcheva, K. (2020*a*), Measuring what counts: The case of rumour stance classification, *in* K.-F. Wong, K. Knight & H. Wu, eds, 'Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing', Association for Computational Linguistics, Suzhou, China, pp. 925–932.
URL: *https://aclanthology.org/2020.aacl-main.92/*

Scarton, C., Silva, D. & Bontcheva, K. (2020*b*), Measuring what counts: The case of rumour stance classification, *in* 'Proceedings of the 1st Conference of the Asia-Pacific Chapter of

the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing', pp. 925–932.

Schick, T. & Schütze, H. (2021), Exploiting cloze-questions for few-shot text classification and natural language inference, *in* P. Merlo, J. Tiedemann & R. Tsarfaty, eds, 'Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume', Association for Computational Linguistics, Online, pp. 255–269.
URL: *https://aclanthology.org/2021.eacl-main.20/*

Schiller, B., Daxenberger, J. & Gurevych, I. (2021), 'Stance detection benchmark: How robust is your stance detection?', *KI-Künstliche Intelligenz* **35**(3), 329–341.

Sennrich, R., Haddow, B. & Birch, A. (2016), Improving neural machine translation models with monolingual data, *in* K. Erk & N. A. Smith, eds, 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Berlin, Germany, pp. 86–96.
URL: *https://aclanthology.org/P16-1009/*

Shi, W., Han, X., Gonen, H., Holtzman, A., Tsvetkov, Y. & Zettlemoyer, L. (2023), Toward human readable prompt tuning: Kubrick's the shining is a good movie, and a good prompt too?, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Findings of the Association for Computational Linguistics: EMNLP 2023', Association for Computational Linguistics, Singapore, pp. 10994–11005.
URL: *https://aclanthology.org/2023.findings-emnlp.733/*

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E. & Singh, S. (2020), AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts, *in* B. Webber, T. Cohn, Y. He & Y. Liu, eds, 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 4222–4235.
URL: *https://aclanthology.org/2020.emnlp-main.346/*

Shu, K., Mahudeswaran, D., Wang, S., Lee, D. & Liu, H. (2020), 'Combating disinformation in a social media age', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **10**(6), e1385.

Shu, R., Nakayama, H. & Cho, K. (2019), Generating diverse translations with sentence codes, *in* A. Korhonen, D. Traum & L. Màrquez, eds, 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 1823–1827.
URL: *https://aclanthology.org/P19-1177/*

Singh, V., Narayan, S., Akhtar, M. S., Ekbal, A. & Bhattacharyya, P. (2017), IITP at SemEval-2017 task 8 : A supervised approach for rumour evaluation, *in* S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer & D. Jurgens, eds, 'Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)', Association for Computational Linguistics, Vancouver, Canada, pp. 497–501.
URL: *https://aclanthology.org/S17-2087/*

Sobhani, P., Inkpen, D. & Zhu, X. (2017), A dataset for multi-target stance detection, *in* M. Lapata, P. Blunsom & A. Koller, eds, 'Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers', Association for Computational Linguistics, Valencia, Spain, pp. 551–557.
URL: *https://aclanthology.org/E17-2088/*

Stacey, J., Cheng, J., Torr, J., Guigue, T., Driesen, J., Coca, A., Gaynor, M. & Johannsen, A. (2024), LUCID: LLM-generated utterances for complex and interesting dialogues, *in* Y. T. Cao, I. Papadimitriou, A. Ovalle, M. Zampieri, F. Ferraro & S. Swayamdipta, eds, 'Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)', Association for Computational Linguistics, Mexico City, Mexico, pp. 56–74.
URL: *https://aclanthology.org/2024.naacl-srw.8/*

Stolfo, A., Wu, B. P., Gurnee, W., Belinkov, Y., Song, X., Sachan, M. & Nanda, N. (2024), Confidence regulation neurons in language models, *in* 'ICML 2024 Workshop on Mechanistic Interpretability'.
URL: *https://openreview.net/forum?id=rB0GsxS5V3*

Su, X., Zhao, Y. & Bethard, S. (2021), The University of Arizona at SemEval-2021 task 10: Applying self-training, active learning and data augmentation to source-free domain adaptation, *in* A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot & X. Zhu, eds, 'Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)', Association for Computational Linguistics, Online, pp. 458–466.
URL: *https://aclanthology.org/2021.semeval-1.56/*

Su, X., Zhao, Y. & Bethard, S. (2022), A comparison of strategies for source-free domain adaptation, *in* S. Muresan, P. Nakov & A. Villavicencio, eds, 'Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Dublin, Ireland, pp. 8352–8367.
URL: *https://aclanthology.org/2022.acl-long.572/*

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. & Hardt, M. (2020), Test-time training with self-supervision for generalization under distribution shifts, *in* 'International conference on machine learning', PMLR, pp. 9229–9248.

Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F. & Wen, J.-R. (2024), Language-specific neurons: The key to multilingual capabilities in large language models, *in* L.-W. Ku, A. Martins & V. Srikumar, eds, 'Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Bangkok, Thailand, pp. 5701–5715.
URL: *https://aclanthology.org/2024.acl-long.309/*

Thorne, J. & Vlachos, A. (2018), Automated fact checking: Task formulations, methods and future directions, *in* E. M. Bender, L. Derczynski & P. Isabelle, eds, 'Proceedings of the 27th International Conference on Computational Linguistics', Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 3346–3359.
URL: *https://aclanthology.org/C18-1283/*

Thorne, J., Vlachos, A., Christodoulopoulos, C. & Mittal, A. (2018), FEVER: a large-scale dataset for fact extraction and VERification, *in* M. Walker, H. Ji & A. Stent, eds, 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)', Association for Computational Linguistics, New Orleans, Louisiana, pp. 809–819.
URL: *https://aclanthology.org/N18-1074/*

Tian, J., Zhang, J., Li, W. & Xu, D. (2021), 'Vdm-da: Virtual domain modeling for source data-free domain adaptation', *IEEE Transactions on Circuits and Systems for Video Technology* **32**(6), 3749–3760.

Tian, Q., Ma, C., Zhang, F.-Y., Peng, S. & Xue, H. (2021), 'Source-free unsupervised domain adaptation with sample transport learning', *Journal of Computer Science and Technology* **36**(3), 606–616.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. et al. (2023), 'Llama 2: Open foundation and fine-tuned chat models', *arXiv preprint arXiv:2307.09288* .

Trung, N. N., Phung, D. & Nguyen, T. H. (2021), Unsupervised domain adaptation for event detection using domain-specific adapters, *in* 'Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021', pp. 4015–4025.

Tzeng, E., Hoffman, J., Saenko, K. & Darrell, T. (2017), Adversarial discriminative domain adaptation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 7167–7176.

Vamvas, J. & Sennrich, R. (2020), X-Stance: A multilingual multi-target dataset for stance detection, *in* 'Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)', Zurich, Switzerland.
URL: *http://ceur-ws.org/Vol-2624/paper9.pdf*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017), 'Attention is all you need', *Advances in neural information processing systems* **30**.

Voita, E., Ferrando, J. & Nalmpantis, C. (2024), Neurons in large language models: Dead, n-gram, positional, *in* L.-W. Ku, A. Martins & V. Srikumar, eds, 'Findings of the Association for Computational Linguistics: ACL 2024', Association for Computational Linguistics, Bangkok, Thailand, pp. 1288–1301.
URL: *https://aclanthology.org/2024.findings-acl.75/*

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A. & Hajishirzi, H. (2020), Fact or fiction: Verifying scientific claims, *in* B. Webber, T. Cohn, Y. He & Y. Liu, eds, 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 7534–7550.
URL: *https://aclanthology.org/2020.emnlp-main.609/*

Wagner, S. S., Behrendt, M., Ziegele, M. & Harmeling, S. (2024), 'The power of llm-generated synthetic data for stance detection in online political discussions', *arXiv preprint arXiv:2406.12480* .

Wang, B., Yin, W., Lin, X. V. & Xiong, C. (2021), Learning to synthesize data for semantic parsing, *in* K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou, eds, 'Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies', Association for Computational Linguistics, Online, pp. 2760–2766.
URL: *https://aclanthology.org/2021.naacl-main.220/*

Wang, L., Li, L., Dai, D., Chen, D., Zhou, H., Meng, F., Zhou, J. & Sun, X. (2023), Label words are anchors: An information flow perspective for understanding in-context learning, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Singapore, pp. 9840–9855.
URL: *https://aclanthology.org/2023.emnlp-main.609/*

Wang, W., Wu, Y., Liu, Y. & Liu, P. (2021), BLCUFIGHT at SemEval-2021 task 10: Novel unsupervised frameworks for source-free domain adaptation, *in* A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot & X. Zhu, eds, 'Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)', Association for Computational Linguistics, Online, pp. 357–363.
URL: *https://aclanthology.org/2021.semeval-1.43/*

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A. S., Arunkumar, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Pal, K. K., Patel, M., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P. R., Verma, P., Puri, R. S., Karia, R., Doshi, S., Sampat, S. K., Mishra, S., Reddy A, S., Patro, S., Dixit, T. & Shen, X. (2022), Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks, *in* Y. Goldberg, Z. Kozareva & Y. Zhang, eds, 'Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 5085–5109.
URL: *https://aclanthology.org/2022.emnlp-main.340/*

Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M. & Le, Q. V. (2022), Finetuned language models are zero-shot learners, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=gEZrGCozdqR*

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V. & Zhou, D. (2024), Chain-of-thought prompting elicits reasoning in large language models, *in* 'Proceedings of the 36th International Conference on Neural Information Processing Systems', NIPS '22, Curran Associates Inc., Red Hook, NY, USA.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. et al. (2022), 'Chain-of-thought prompting elicits reasoning in large language models', *Advances in neural information processing systems* **35**, 24824–24837.

Wei, J. & Zou, K. (2019), EDA: Easy data augmentation techniques for boosting performance on text classification tasks, *in* K. Inui, J. Jiang, V. Ng & X. Wan, eds, 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 6382–6388.
URL: *https://aclanthology.org/D19-1670/*

Wei, P., Xu, N. & Mao, W. (2019), Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity, *in* K. Inui, J. Jiang, V. Ng & X. Wan, eds, 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 4787–4798.
URL: *https://aclanthology.org/D19-1485/*

Wilby, D., Karmakharm, T., Roberts, I., Song, X. & Bontcheva, K. (2023), GATE teamware 2: An open-source tool for collaborative document classification annotation, *in* D. Croce & L. Soldaini, eds, 'Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations', Association for Computational Linguistics, Dubrovnik, Croatia, pp. 145–151.
URL: *https://aclanthology.org/2023.eacl-demo.17/*

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. (2020), Transformers: State-of-the-art natural language processing, *in* Q. Liu & D. Schlangen, eds, 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', Association for Computational Linguistics, Online, pp. 38–45.
URL: *https://aclanthology.org/2020.emnlp-demos.6/*

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. et al. (2019), 'Huggingface's transformers: State-of-the-art natural language processing', *arXiv preprint arXiv:1910.03771* .

Wu, B., Li, Y., Mu, Y., Scarton, C., Bontcheva, K. & Song, X. (2023*a*), Don't waste a single annotation: improving single-label classifiers through soft labels, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Findings of the Association for Computational Linguistics: EMNLP 2023', Association for Computational Linguistics, Singapore, pp. 5347–5355.
URL: *https://aclanthology.org/2023.findings-emnlp.355/*

Wu, B., Li, Y., Mu, Y., Scarton, C., Bontcheva, K. & Song, X. (2023*b*), Don't waste a single annotation: improving single-label classifiers through soft labels, *in* 'Findings of the Association for Computational Linguistics: EMNLP 2023', pp. 5347–5355.
URL: *https://aclanthology.org/2023.findings-emnlp.355/*

Wu, X., Yao, W., Chen, J., Pan, X., Wang, X., Liu, N. & Yu, D. (2024), From language modeling to instruction following: Understanding the behavior shift in LLMs after instruction tuning, *in* K. Duh, H. Gomez & S. Bethard, eds, 'Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)', Association for Computational Linguistics, Mexico City, Mexico, pp. 2341–2369.
URL: *https://aclanthology.org/2024.naacl-long.130/*

Xu, C., Paris, C., Nepal, S. & Sparks, R. (2018), Cross-target stance classification with self-attention networks, *in* I. Gurevych & Y. Miyao, eds, 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)', Association for Computational Linguistics, Melbourne, Australia, pp. 778–783.
URL: *https://aclanthology.org/P18-2123/*

Xu, Z., Cohen, D., Wang, B. & Srikumar, V. (2024), In-context example ordering guided by label distributions, *in* K. Duh, H. Gomez & S. Bethard, eds, 'Findings of the Association for Computational Linguistics: NAACL 2024', Association for Computational Linguistics, Mexico City, Mexico, pp. 2623–2640.
URL: *https://aclanthology.org/2024.findings-naacl.167/*

Yang, C., Guo, X., Chen, Z. & Yuan, Y. (2022), 'Source free domain adaptation for medical image segmentation with fourier style mining', *Medical Image Analysis* **79**, 102457.

Yang, R., Xie, W., Liu, C. & Yu, D. (2019), BLCU_NLP at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation, *in* J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki & S. M. Mohammad, eds, 'Proceedings of the 13th International Workshop on Semantic Evaluation', Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 1090–1096.
URL: *https://aclanthology.org/S19-2191/*

Yang, Z., Dai, D., Wang, P. & Sui, Z. (2023), Not all demonstration examples are equally beneficial: Reweighting demonstration examples for in-context learning, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Findings of the Association for Computational Linguistics: EMNLP 2023', Association for Computational Linguistics, Singapore, pp. 13209–13221.
URL: *https://aclanthology.org/2023.findings-emnlp.880/*

Yarowsky, D. (1995), Unsupervised word sense disambiguation rivaling supervised methods, *in* '33rd Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Cambridge, Massachusetts, USA, pp. 189–196.
URL: *https://aclanthology.org/P95-1026/*

Yoo, K. M., Kim, J., Kim, H. J., Cho, H., Jo, H., Lee, S.-W., Lee, S.-g. & Kim, T. (2022), Ground-truth labels matter: A deeper look into input-label demonstrations, *in* Y. Goldberg, Z. Kozareva & Y. Zhang, eds, 'Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 2422–2437.
URL: *https://aclanthology.org/2022.emnlp-main.155/*

Yoon, S., Kim, Y. & Jung, K. (2021), Self-adapter at SemEval-2021 task 10: Entropy-based pseudo-labeler for source-free domain adaptation, *in* A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot & X. Zhu, eds, 'Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)', Association for Computational Linguistics, Online, pp. 452–457.
URL: *https://aclanthology.org/2021.semeval-1.55/*

Yu, J., Jiang, J., Khoo, L. M. S., Chieu, H. L. & Xia, R. (2020), Coupled hierarchical transformer for stance-aware rumor verification in social media conversations, *in* B. Webber, T. Cohn, Y. He & Y. Liu, eds, 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 1392–1401.
URL: *https://aclanthology.org/2020.emnlp-main.108/*

Zhang, R., Tian, Y., Wei, P., Zeng, D. D. & Mao, W. (2024), An LLM-enabled knowledge elicitation and retrieval framework for zero-shot cross-lingual stance identification, *in* Y. Al-Onaizan, M. Bansal & Y.-N. Chen, eds, 'Findings of the Association for Computational Linguistics: EMNLP 2024', Association for Computational Linguistics, Miami, Florida, USA, pp. 12253–12266.
URL: *https://aclanthology.org/2024.findings-emnlp.714/*

Zhang, R., Yang, H. & Mao, W. (2023), Cross-lingual cross-target stance detection with dual knowledge distillation framework, *in* 'Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing', pp. 10804–10819.

Zhang, X., Li, S., Hauer, B., Shi, N. & Kondrak, G. (2023), Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs, *in* H. Bouamor, J. Pino & K. Bali, eds, 'Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Singapore, pp. 7915–7927.
URL: *https://aclanthology.org/2023.emnlp-main.491/*

Zhang, X., Zhao, J. & LeCun, Y. (2015), 'Character-level convolutional networks for text classification', *Advances in neural information processing systems* **28**.

Zhang, Y., Feng, S. & Tan, C. (2022), Active example selection for in-context learning, *in* Y. Goldberg, Z. Kozareva & Y. Zhang, eds, 'Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 9134–9148.
URL: *https://aclanthology.org/2022.emnlp-main.622/*

Zhang, Z., Li, Y., Zhang, J. & Xu, H. (2024), LLM-driven knowledge injection advances zero-shot and cross-target stance detection, *in* K. Duh, H. Gomez & S. Bethard, eds, 'Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)', Association for Computational Linguistics, Mexico City, Mexico, pp. 371–378.
URL: *https://aclanthology.org/2024.naacl-short.32/*

Zhang, Z., Zhang, A., Li, M. & Smola, A. (2023), Automatic chain of thought prompting in large language models, *in* 'The Eleventh International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=5NTt8GFjUHkr*

Zhao, Z., Ma, Z., Lin, Z., Xie, J., Li, Y. & Shen, Y. (2024), Source-free domain adaptation for aspect-based sentiment analysis, *in* N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue, eds, 'Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)', ELRA and ICCL, Torino, Italia, pp. 15076–15086.
URL: *https://aclanthology.org/2024.lrec-main.1310/*

Zheng, J., Baheti, A., Naous, T., Xu, W. & Ritter, A. (2022), Stanceosaurus: Classifying stance towards multicultural misinformation, *in* Y. Goldberg, Z. Kozareva & Y. Zhang, eds, 'Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 2132–2151.
URL: *https://aclanthology.org/2022.emnlp-main.138/*

Zhou, W., Ge, T., Xu, K., Wei, F. & Zhou, M. (2019), BERT-based lexical substitution, *in* A. Korhonen, D. Traum & L. Màrquez, eds, 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 3368–3373.
URL: *https://aclanthology.org/P19-1328/*

Zhu, Z., Zhang, Z., Zhang, H. & Li, C. (2025), RATSD: Retrieval augmented truthfulness stance detection from social media posts toward factual claims, *in* L. Chiruzzo, A. Ritter & L. Wang, eds, 'Findings of the Association for Computational Linguistics: NAACL 2025', Association for Computational Linguistics, Albuquerque, New Mexico, pp. 3366–3381.
URL: *https://aclanthology.org/2025.findings-naacl.187/*

Zotova, E., Agerri, R., Nuñez, M. & Rigau, G. (2020), Multilingual stance detection in tweets: The Catalonia independence corpus, *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis, eds, 'Proceedings of the Twelfth Language Resources and Evaluation Conference', European Language Resources Association, Marseille, France, pp. 1368–1375.
URL: *https://aclanthology.org/2020.lrec-1.171/*

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. & Procter, R. (2018*a*), 'Detection and resolution of rumours in social media: A survey', *ACM Computing Surveys (CSUR)* **51**(2), 1–36.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. & Procter, R. (2018*b*), 'Detection and resolution of rumours in social media: A survey', *ACM Comput. Surv.* **51**(2).

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. & Procter, R. (2018*c*), 'Detection and resolution of rumours in social media: A survey', *ACM Comput. Surv.* **51**(2).
URL: *https://doi.org/10.1145/3161603*

Zubiaga, A., Liakata, M., Procter, R., Bontcheva, K. & Tolmie, P. (2015), Crowdsourcing the annotation of rumourous conversations in social media, *in* 'Proceedings of the 24th international conference on World Wide Web', pp. 347–353.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G. & Tolmie, P. (2016), 'Analysing how people orient to and spread rumours in social media by looking at conversational threads', *PloS one* **11**(3), e0150989.