Modelling Social Media Popularity of News Articles Using Headline Text

Alicja Piotrkowicz

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

The University of Leeds
School of Computing
December 2017

Declarations

The candidate confirms that the work submitted is her own, except where work which has formed part of a jointly authored publication has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Some parts of the work presented in this thesis have been published in the following articles:

Piotrkowicz, A., Dimitrova, V.G. and Markert, K. (2016) Automatic Extraction of News Values from Headline Text. In: Proceedings of EACL. European Chapter of the Association for Computational Linguistics, 03-07 Apr 2017, Valencia, Spain.

Piotrkowicz, A., Dimitrova, V.G., Otterbacher, J. and Markert, K. (2017) Headlines Matter: Using Headlines to Predict the Popularity of News Articles on Twitter and Facebook. In: Proceedings of ICWSM, Short Papers. International Conference of Web and Social Media, 15-18 May 2017, Montreal, Canada.

Piotrkowicz, A., Dimitrova, V.G., Otterbacher, J. and Markert, K. (2017) The Impact of News Values and Linguistic Style on the Popularity of Headlines on Twitter and Facebook. In: Proceedings of the Second International Workshop on News and Public Opinion (ICWSM NECO 2017). International Conference of Web and Social Media, 15-18 May 2017, Montreal, Canada.

Dimitrova, V.G., Mitrovic, A., **Piotrkowicz, A.**, Lau, L., and Weerasinghe, A. (2017) Using Learning Analytics to Devise Interactive Personalised Nudges for Active Video Watching. In: Proceedings of 25th Conference on User Modeling, Adaptation and Personalization. UMAP2017 – 25th Conference on User Modeling, Adaptation and Personalization, 09-12 Jul 2017, Bratislava, Slovakia. Association for Computing Machinery.

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

Acknowledgements

I would like to extend my heartfelt thanks to my supervisor, Dr Vania Dimitrova. Her guidance and support have been invaluable in making this thesis what it is today. I am infinitely grateful for all the time that she has given to this thesis and my other research activities. I have learnt an enourmous amount from her and thanks to that I have grown not only as a researcher, but as a person.

My warmest thanks go to Prof Katja Market, my supervisor at the beginning of this PhD. It is thanks to her that I embarked on this PhD journey and her encouragements were invaluable during the first year of this PhD. I remember very fondly our early discussions which shaped this project.

I would also like to thank Dr Justin Washtell who was my advisor for this thesis. His advice has always been very valuable and strengthened the methodology of this thesis.

A very warm thanks to my fellow PhD students in the School of Computing, and particularly to Sam, Matt, Leroy, Luke, Marwan, Entisar, Noushin, Aryana, Muhannad, Paul, and Jawad. Thanks to you this PhD experience was not just an academic one.

To my friends, many thanks for supporting me these past few years and listening to my PhD-related grumbles. I'm particularly grateful to Marjorie, Diletta, and Caterina for being the best cheerleaders I could hope for. Special thanks go to Toby whose support kept me going in the final write-up period.

Moje najszczersze podziękowania kieruję do moich Rodziców i Siostry. To przede wszystkim dzięki Ich wsparciu byłam w stanie tyle osiągnąć.

Abstract

The way we formulate headlines matters – this is the central tenet of this thesis.

Headlines play a key role in attracting and engaging online audiences. With the increasing usage of mobile apps and social media to consume news, headlines are the most prominent – and often the only – part of the news article visible to readers. Earlier studies examined how readers' preferences and their social network influence which headlines are clicked or shared on social media. However, there is limited research on the impact of the headline text on social media popularity.

To address this research gap we pose the following question: how to formulate a headline so that it reaches as many readers as possible on social media. To answer this question we adopt an experimental approach to model and predict the popularity of news articles on social media using headlines. First, we develop computational methods for an automatic extraction of two types of headline characteristics. The first type is news values: Prominence, Sentiment, Magnitude, Proximity, Surprise, and Uniqueness. The second type is linguistic style: Brevity, Simplicity, Unambiguity, Punctuation, Nouns, Verbs, and Adverbs. We then investigate the impact of these features on popularity using social media popularity on Twitter and Facebook, and perceived popularity obtained from a crowdsourced survey. Finally, using these features and headline metadata we build prediction models for global and country-specific social media popularity. For the country-specific prediction model we augment several news values features with country relatedness information using knowledge graphs.

Our research established that computational methods can be reliably used to characterise headlines in terms of news values and linguistic style features; and that most of these features significantly correlate with social media popularity and to a lesser extent with perceived popularity. Our prediction model for global social media popularity outperformed state-of-the-art baselines, showing that headline wording has an effect on social media popularity. With the country-specific prediction model we showed that we improved the features implementations by adding data from knowledge graphs.

These findings indicate that formulating a headline in a certain way can lead to wider readership engagement. Furthermore, our methods can be applied to other types of digital content similar to headlines, such as titles for blog posts or videos. More broadly our results signify the importance of content analysis for popularity prediction.

Contents

List of Tables

List of Figures

Abbreviations and Symbols

API Application Programming Interface

GPS Global Positioning System

IAA Inter-Annotator Agreement

IR Information Retrieval

LL Log-likelihood

MAE Mean Absolute Error

NLP Natural Language Processing

NLTK Natural Language Toolkit

NP Noun Phrase

POS Part of Speech

RDF Resource Description Framework

RBF Radial Basis Function

SD Standard Deviation

SVM Support Vector Machine

URL Unique Resource Locator

VP Verb Phrase

C Set of content words (nouns, verbs, adjectives, or adverbs) in a headline

e Entity in a headline

E Set of entities in a headline

F news article's Facebook popularity after three days

F_1 Harmonic mean of precision and recall

H Set of tokens in a headline

κ Fleiss' kappa

$\mathcal{M}$ Our full feature model

$\mathcal{M}_{\mathcal{A}}$ Reimplementation of Arapakis et al. (2014) baseline

$\mathcal{M}_{\mathcal{B}}$ Reimplementation of Bandari et al. (2012) baseline

$\mathcal{M}_{\mathcal{M}}$ Model with metadata features only

$\mathcal{M}_{\mathcal{N}}$ Model with news values features only

$\mathcal{M}_{\mathcal{N}+\mathcal{S}}$ Model with news values and style features only

$\mathcal{M}_{\mathcal{S}}$ Model with linguistic style features only

$\mathcal{M}_{\mathcal{U}}$ Unigrams baseline

R^2 Coefficient of determination

ρ Spearman's rank correlation coefficient

τ Kendall rank correlation coefficient

T news article's Twitter popularity after three days

Conventions

In this thesis when discussing our model *social media popularity* refers to *popularity on Twitter and Facebook*. *Popularity of a headline* refers to the *popularity of the news article featuring that headline*.

The feature group *linguistic style* is sometimes abbreviated to *style*.

We refer to the *popularity obtained using responses from the crowdsourced survey* as *perceived popularity*.

TagMe entities are written in small caps, e.g. UNITED KINGDOM.

Throughout this thesis *we* refers to the *author* and *ours* refers to the *author's*.

Chapter 1

Introduction

The digital landscape that we inhabit offers limitless distractions. Both the providers and readers of online content need a way of navigating that environment. One of the key signposting methods that is used are *headlines*. Headlines are meant to catch attention and direct online readers to the most relevant, interesting, or profitable content. The fundamental issue for people who write headlines is *how to more effectively attract readers*. This question applies not only to journalists (who routinely write headlines and have received appropriate training), but to all authors of the myriads of user-generated content from blog posts to videos, which feature a title.

This thesis proposes and evaluates a solution – quantifying news values and linguistic style of headlines in order to model their popularity on social media. To scope the research we focus on broadsheet news articles and explore their popularity on Twitter and Facebook, since social media networks have become an integral part of the news cycle. While there have been computational studies of linguistic styles' impact on online content, these have been applied only to a limited extent to headlines and we are the first to propose a fully automatic operationalisation of news values (aspects which are said to influence newsworthiness of news stories according to journalism studies literature) from headline text. We evaluate and apply these methods in two types of experimental settings. Firstly, we calculate correlations and conduct a crowdsourced survey to investigate the impact of individual features on social media popularity, thus gaining insight into how a headline can be reformulated to achieve higher popularity. Secondly, we build global and country-specifc prediction models in order to gain an expectation of a response on social media, which can prompt the reformulation of a headline. The country-specific prediction models make use of reimplementations of news values which combine natural processing methods with semantic information from Wikidata, thus obtaining a new state-of-the-art for these implementations. The news values operationalisation is useful not just for our task of modelling social media popularity, but also in a number of other domains like recommender systems, writing support, or discourse analysis.

1.1 Motivation

There are several factors which influenced the research focus chosen for this project.

Firstly, for the online environment in general and social media platforms in particular, **headlines** play a very prominent role. Headlines are usually the first thing a reader notices on a news website, and sometimes they are their only introduction to an article. Furthermore, when an article is shared on social media, often one can only see the headline (e.g. when retweeting a news article or sharing it on Facebook). It has also been noted that apart from just aiming to catch attention, headlines now often play the role of summaries. Gabielkov et al. (2016) found that 59% of the shared URLs pointing to news content are never clicked (i.e. shared without accessing the content). If headlines are frequently treated as summaries, then that might have a bearing on what kind of phrasing is preferred. For example, a surprising or funny headline might catch the reader's attention, but if they are looking for an at-a-glance summary of the daily news, then perhaps that is not the preferred wording. With the key role that headlines play in the online environment, we need to understand what impact headline phrasing has on popularity. Our work explores a variety of textual factors relating to headline phrasing and their effect on social media popularity.

Secondly, social media networks have become an integral part of the news environment. Not only is news content being increasingly disseminated through social media, but social media allows ordinary Internet users to engage in the news production process. This can happen either by users supplying event information or writing their own news stories (citizen journalism), or more commonly through curating content by sharing on social media what they deem newsworthy. The very disruptive consequences of social media entering (and taking over) the news domain have been highlighted by the Columbia Journalism Review:

> "Our news ecosystem has changed more dramatically in the past five years than perhaps at any time in the past five hundred. [...] Social media hasn't just swallowed journalism, it has swallowed everything."[1]

[1] https://www.cjr.org/analysis/facebook_and_media.php [Accessed 13th April 2018]

One of the solutions proposed (although not without its own dangers) is for news outlets to engage to a greater extent with their readers through social media. Another institution, Pew Research Center, stresses the news dissemination aspect of social media and how that can influence news outlets' news production practices:

> "Understanding not only what content users will want to consume but also what content they are likely to pass along may be a key to how stories are put together and even what stories get covered in the first place.". (Olmstead et al., 2011)

In order to engage more effectively with readers on social media platforms, news outlets need to know: (i) what aspects of the news content – and particularly headlines – are popular among social media users (which impacts news selection), (ii) how to phrase their content – including headlines – to attract a wider audience (which impacts content production). Our work investigates specifically the wording of headline text and its effect on social media popularity.

Thirdly, in terms of modelling popularity there is a significant advantage to using **features derived from the text**, namely that the text is available prior to the release of the content online. Using only pre-publication data allows the user (for example, a journalist) to have an expectation of the social media response prior to publishing the text. This allows the author to try out different versions of the same text before publication. In many domains such as news, marketing, or public relations, this is a significant advantage. Furthermore, there are also cases where only the text is available, with no interaction or early adopter data. For example, the 'cold-start' problem in recommender systems refers to a scenario when a new user enters a system and there is no user model available for them. In that scenario, predicting the popularity of a piece of content (e.g. a news article) from the text of the headline would be very beneficial.

Finally, the **automatic extraction of news values and style features** from headlines can be a central tool for a range of applications. Headline newsworthiness insights would be directly beneficial to news outlets trying to engage with social media users. They can also be incorporated more widely into online multimedia content publishing, e.g. YouTube[2], and writing support software, e.g. Scrivener[3] or Hemingway[4]. In these systems insights about headline wording (based on the correlations of news values and linguistic style features with social media popularity) can be used to guide authors on how to compose or

[2] https://www.youtube.com/ [Accessed 13th April 2018]

[3] http://www.literatureandlatte.com/scrivener.php [Accessed 13th April 2018]

[4] http://www.hemingwayapp.com/ [Accessed 13th April 2018]

reformulate the headline text to attract audiences' attention. Furthermore, computational methods of deriving news values at scale can help digital humanities researchers conduct large-scale comparisons of news values across digital outlet types, genres, demographics, etc. These can be complementary to traditionally used qualitative studies.

1.2 Scope

In this thesis we model the social media popularity of news articles using headlines. Below we discuss several factors which delimit the scope of our investigation.

News corpora. We develop and evaluate our methods using headlines corpora obtained from news outlets that are representative of a wide range of news publications under the umbrella of 'broadsheet' or 'quality', as opposed to tabloid newspapers which differ in style and tone. We chose broadsheet news sources, because many NLP tools have been developed and trained on newswire corpora which consist of broadsheet news outlets like *New York Times* and *Associated Press*. In this thesis we use headlines from *The Guardian* and *New York Times*. They are both broadsheet news sources, but they differ in writing style and coverage, which helps us to understand the generalisability of our methods.

Popularity measures. In this thesis we focus on social media popularity, which we define as the *amount of social media attention*. In particular we use tweets and retweets from Twitter, and likes and shares from Facebook. This decision is motivated by the special role these two news websites play in disseminating news content (cf. Section 2.2.1). We do not consider secondary metrics of popularity, such as the number comments on Facebook, as this could introduce a degree of noise (e.g. a person responding to their friend rather than reacting to the news article).

Feature engineering for news values. Our methods for operationalising news values rely on how they are realised through explicit linguistic indicators in headline text. This decision relates to our main hypothesis (discussed in detail in the next section) the formulation of a headline influences its popularity on social media. By investigating explicit linguistic indicators we can make recommendations on how to reformulate a headline, so that it reaches higher social media popularity. Furthermore, we make the implementations as generic and domain-independent as possible. That is to say, although we are using news corpora, we want our methods to be applicable to other domains.

External knowledge sources. As headline text does not provide much contextual information, we enrich it by making use of external knowledge resources. We adopt an entity-driven approach, whereby entity mentions in the headline text are related to a knowledge model. Specifically, we use Wikipedia and the knowledge graph behind it, Wikidata.

Currently English Wikipedia consists of over 5.5 million articles and approximately 600 are added every day[5]. It provides a large-scale, generic resource which can be computationally accessed and queried using Wikidata.

1.3 Hypothesis and Research Questions

The core hypothesis of this thesis is that *the way that a headline is formulated has an impact on the social media popularity of the news article*.

Within that core hypothesis there are several research questions which we address in this thesis.

RQ1: Can news values be reliably extracted from headline text?

RQ2: What is the impact of headline-derived news values and style features on social media popularity?

RQ3: What is the impact of headline-derived news values and style features on perceived popularity and how is it judged by readers?

RQ4: To what extent can headline-derived news values and style features be used to predict the social media popularity of news articles?

RQ5: Does augmenting the feature engineering with country-specific information improve the impact of that feature on social media popularity?

We address each of these questions in the following chapters of this thesis: RQ1 in Chapter 4, RQ2 in Chapter 6, RQ3 in Chapter 7, RQ4 in Chapter 8, and finally RQ5 in Chapter 9.

1.4 Methods

The methods used in this thesis for the task of modelling social media popularity of news articles using headlines can be separated into the following steps described below.

1. Data Collection and Preprocessing

   (a) News headlines

   (b) Social media popularity

2. Feature Extraction

[5] https://en.wikipedia.org/wiki/Wikipedia:Statistics [Accessed 13th April 2018]

   (a) News values
   (b) Linguistic style

3. Impact of Features
   (a) Social media popularity
   (b) Perceived popularity

4. Prediction Model
   (a) Global social media popularity
   (b) Country-specific social media popularity

The first step in our methodology is data collection and preprocessing. We first collect headlines from two broadsheet news outlets: *The Guardian* and *New York Times*. We then obtain social media popularity measures for each headline, which will later be used as target variables in correlations studies and prediction models. The headlines are preprocessed to obtain a set of tokens, content words, and named entities, which form the basis of further feature engineering.

In the second step we extract two types of features from headline text: news values and linguistic style. These two feature groups cover a variety of headline aspects and provide two complementary perspectives on social media popularity of headlines: journalistic perspective through news values, and advice on wording through linguistic style.

We evaluate the features we propose in two ways. In the third step of our methodology we investigate the correlations of individual news values and linguistic style features with social media popularity and with perceived popularity obtained through a crowdsourced survey. This gives us insight into how a headline can be reformulated to achieve higher popularity on social media. Then in the fourth step we use these features to predict global and country-specific popularity. The prediction models give us an indication of the expected response on social media for a particular headline, which can prompt the reformulation of the headline if the predicted popularity is not satisfactory. The country-specific prediction model allows us to further investigate several of our proposed features by reimplementing them using semantic resources and user demographics.

1.5 Contributions

The work presented in this thesis makes the following contributions:

(i) created headlines corpora with associated social media popularity measures for two news outlets;

(ii) developed and evaluated computational methods for automatically extracting six news values from headline text;

(iii) investigated the impact of headline-derived news values and style features on social media popularity and on perceived popularity obtained from a crowdsourced survey;

(iv) built and evaluated prediction models which used news values and style features, including general and country-specific popularity.

1.6 Thesis Structure

The thesis is structured as follows.

Chapter 2: Related Work presents an overview of the relevant literature. It summarises the earlier work on online news article popularity prediction and categorises it according to several factors. We then turn to overviews of research which directly influence our approach. Starting with the importance of headlines as part of the digital experience, through the effect of wording on online popularity, to the journalistic perspective on our task encapsulated by the concept of news values.

Chapter 3: Data Collection and Preprocessing describes the data collection process for the headlines corpora. We use two sources: *The Guardian* and *New York Times*. The chapter gives an overview of the datasets. We then detail the process of obtaining and validating the popularity metrics on Twitter and Facebook for the collected articles.

Chapter 4: Implementing News Values presents the description and implementation of six news values (Prominence, Sentiment, Magnitude, Proximity, Surprise, Uniqueness), and provides the results of applying these methods to the headlines corpora. The proposed news values operationalisation is evaluated against a manually annotated gold standard.

Chapter 5: Implementing Linguistic Style describes the implementation of linguistic style features and the application of these methods on the headlines corpora. For this we have utilised state-of-the-art natural language processing tools.

Chapter 6: Impact of News Values and Style on Social Media Popularity reports and discusses the results of correlating headline-derived news values and style feature values with news article popularity on social media. This allows us to examine each feature individually and to show its correlation with social media popularity.

Chapter 7: Impact of News Values and Style on Personal Engagement provides a complementary perspective on the impact of news values and style on perceived popularity obtained using a crowdsourced survey. We report a correlations study, qualitative analysis of survey responses by expert annotators, and judgements about news values and style impact by survey participants.

Chapter 8: Social Media Popularity Prediction Using News Values and Style brings together news values and style features derived from headlines and uses them in a prediction model of news article popularity on social media. Our prediction model is compared against several baselines. We also conduct two further investigations: one which uses feature subsets, and one which uses corpus subsets.

Chapter 9: Country-specific Prediction Model outlines the process of running a prediction model which takes into account the reader location. We first describe the methods of gathering location data and the resulting dataset. We reimplement two news values features – Prominence and Proximity – using geographic relevance data obtained from Wikidata. To judge the effectiveness of the retuning method we conduct a correlations study and run a prediction model comparing it to a location-naive model.

Chapter 10: Conclusions and Future Work summarises our findings and offers some suggestions for future work.

Chapter 2

Related Work

Our task of modelling the social media popularity of news articles using headline text touches upon a number of domains. We start with a broader view of modelling popularity (and in particular predicting popularity) of various types of online content in Section 2.1. Then in Section 2.2 we review in detail the research on predicting the particular type of content that we focus on in this thesis: news articles. The next three sections present the literature that motivates the various aspects of our approach which we introduce or further develop for the task of modelling popularity of news articles. Our decision to focus on headlines is motivated in Section 2.3 which looks at the importance of headlines and the challenges in processing headline text. As headlines are a type of short text, Section 2.4 reviews the research on the effect of wording on short text popularity. Finally, as we are working on corpora from the news domain in Section 2.5 we present the literature on news values, which offer a journalistic perspective on our task of modelling social media popularity of news articles using headlines.

2.1 Popularity on Social Media

Before we present the literature about the specific task we are attempting (i.e. modelling the social media popularity of news articles) we take a broader view on the task of predicting the popularity of online content on social media.

2.1.1 Availability of Online Content

Webster (2014) observed that “the widening gap between limitless media and limited attention makes it a challenge for anything to attract an audience”. Indeed, online audiences are now faced with a vast amount of varied content. Some examples include: news (e.g. Chartbeat reports over 92,000 articles published online daily[1]), blogs (e.g. nearly 70 million Wordpress posts published each month[2]), forum posts (e.g. over 73 million Reddit submissions in 2015[3]), and videos (e.g. 300 hours of video uploaded every minute on YouTube[4]). This online content is not just produced, but also engaged with at a massive scale. For example, a recent study by AOL Nielsen[5] reports that 27 million pieces of online content are shared daily in the United States. For content creators, the challenge is to attract audiences to engage with the content they have authored.

Online content production varies. Many content creators are professional. One of the major professional online content creators are news outlets. They also produce a considerable amount of content; for example, 500 daily articles from *The Washington Post*, 240 from *Wall Street Journal*, and 230 from *New York Times*[6]. However, much of the content produced online is created by ordinary Internet users – the so-called user-generated content. In August 2012, 46% of American adult Internet users posted original content online[7]. Within the news domain, citizen journalism (i.e. ordinary citizens creating or co-creating news) has been the subject of much debate (e.g. Wall (2015)). With every Internet user being a potential creator of content – leading to the production of vast amounts of online content – computational approaches are needed to understand what attracts the online audiences’ attention. Thus, research on online content popularity prediction can inform the practicces of both content creators and consumers. We give examples of online content popularity prediction in the following section.

2.1.2 Social Media and Predicting Popularity

Social media websites are very widely used. In the United States 65% of all adults (76% of Internet users) use social media websites (Perrin, 2015). This figure rises to 90% for all adults aged 18-29. Among the most popular social networking sites were: Facebook (71% of American online adults), LinkedIn (28%), Pinterest (28%), Instagram (26%), and

[1] http://bit.ly/2cPaQZK [Accessed 13th April 2018]

[2] https://wordpress.com/activity/ [Accessed 13th April 2018]

[3] https://redditblog.com/2015/12/31/reddit-in-2015/ [Accessed 13th April 2018]

[4] http://www.statisticbrain.com/youtube-statistics/ [Accessed 13th April 2018]

[5] https://bit.ly/2HqTYWP [Accessed 13th April 2018]

[6] http://www.theatlantic.com/technology/archive/2016/05/how-many-stories-do-newspapers-publish-per-day/483845/ [Accessed 13th April 2018]

[7] http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/ [Accessed 13th April 2018]

Twitter (23%) (Duggan et al., 2015). For many people checking social media websites is part of their daily routine with 70% of Facebook users and 36% of Twitter users checking the site daily.

Many researchers have used social media popularity as an approximation of popularity overall – a 'sensor' of public opinion. Social media data was used to predict a variety of outcomes with varying degrees of success. Asur and Huberman (2010) used Twitter data (tweet rate and tweet polarity) to predict box office revenues. Tumasjan et al. (2010) used the number of tweets mentioning a political party as a potential reflection of the vote share. Stewart et al. (2012) used keyword volume on Twitter to forecast daily consumer spending. Bollen et al. (2011) tried using public sentiment on Twitter to improve the performance of a stock market prediction model. In many cases Twitter data contributed to prediction performance, suggesting that it could be used to capture public opinion. This makes it a valuable target for measuring popularity of online content.

What follows is the importance of predicting the popularity of content on social media, such as Twitter or YouTube. Social media popularity is related to the concept of engagement on social media. Social media engagement encompasses a wide variety of actions, such as replying, sharing, favouriting (e.g. Rowe and Alani (2014) who used social and content features to predict replying behaviour across five social media platforms). There are considerable disparities in defining social media popularity, however quite often popularity is defined as the amount of attention (e.g. likes, tweets), but not necessarily any deeper interaction (e.g. commenting, replying). We define popularity as *the amount of attention* that a news article gets on social media, as measured by the number of tweets and retweets on Twitter, and likes and shares on Facebook. A variety of approaches have been used to predict popularity of social media content. Szabo and Huberman (2010) used early viewing/voting patterns on YouTube and Digg to predict the long-term popularity of that content. Similarly, Ahmed et al. (2013) used temporal patterns to predict popularity of content on YouTube, Digg, and Vimeo. Video content (as in topic) is challenging to operationsalise as features, which might explain the usage of temporal patterns for prediction. However, Figueiredo et al. (2014) found through a crowdsourcing task that the content preferred by users reached higher popularity on YouTube, highlighting the importance of considering content for online popularity prediction. Predicting the popularity of Twitter content like tweets and hashtags has also been the subject of research. Petrovic et al. (2011) used social (e.g. number of followers) and content (e.g. number of hashtags) features to predict whether a tweet would be retweeted. They found social features to be better predictors. Artzi et al. (2012) built a classifier for predicting whether a tweet will get a response. They used six feature categories: historical (e.g. ratio of tweets by user that were retweeted), social (e.g. number of followers), aggregate lexical (e.g. ratio of response to non-response tweets for bigrams), local content (e.g. number of hashtags), posting (e.g. time and day of the week), and sentiment (e.g. number of positive and negative sentiment words in a tweet) features. They found that historical, social, and aggregate lexical features contributed most to classification performance. Kong et al. (2014) looked at hashtags and their lifecycle (whether a hashtag will become viral and for how long). They used the following feature categories: meme (e.g. tweet count), user (e.g. total follower count), content (e.g. emoticon count), network (e.g. graph density), hashtag (e.g. hashtag length), time series (e.g. dormant period), and prototype (i.e. similarity to historical hashtags). They found that time series features contribute the most to prediction performance overall, while the predictive power of some feature types depended on the timing of prediction (e.g. prototype features worked well shortly after a hashtag had become active, while user and content features worked better at later time points).

Overwhelmingly social media popularity has been estimated using Twitter, and to a lesser extent YouTube. Although Facebook has by far the highest number of users of any social media website[8], it is not commonly used for popularity prediction. That is because most Facebook data is private and explicit consent from the user needs to be obtained in most cases. The small amount of research on predicting popularity on Facebook focuses on marketing pages. Yu and Kak (2012) used keywords in Facebook messages posted by restaurants to predict whether they will be popular (measured by the number of likes). They found that with this approach they could only significantly outperform the baseline when looking at a subset of most and least popular messages, but not when looking at the whole dataset. Lakkaraju and Ajmera (2011) found that using content features (e.g. sentiment, brand-related keywords) and some author and temporal features outperforms a user-centric baseline in predicting the number of comments on a brand post on Facebook.

Unlike videos, for Twitter and Facebook messages content features are readily available and commonly used in prediction models. However, they are not always found to be the best predictors. While social features like follower count have repeatedly been shown to be good predictors of Twitter popularity, this is not useful from the perspective of content creator, as the number of followers cannot be easily changed. On the other hand, content features lend themselves to editing, which is why investigating the predictive power of a wide variety of content-derived features can help content authors. Furthermore, Zhang et al. (2014) looked at tweet popularity from an information processing perspective and found that content factors carry more weight than contextual factors in explaining popularity.

[8] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ [Accessed 13th April 2018]

They found tweet topic and affect to have most significant impact. Although we are working on headlines rather than the types of content covered in this section, we implement content features which were found to be good predictors of social media popularity: Uniqueness (similar to prototype features in Kong et al. (2014)) and Sentiment (similar to affect in Zhang et al. (2014) and sentiment in Lakkaraju and Ajmera (2011)). We add to previous work by considering a wide range of content features (cf. journalism-inspired news values in Chapter 4 and linguistic style in Chapter 5), and by using the same set of features to predict popularity on both Twitter and Facebook which allows the comparison between the two platforms.

In the following section we review in more detail the research on popularity prediction of news articles – the online content that we focus on in this thesis.

2.2 News Article Popularity Prediction

In this thesis we focus on the social media popularity of content from the news domain. In this section we first give an overview of social media as a vehicle of news dissemination, which motivates our choice of social media popularity as the target metric. Then we summarise the research on news article popularity prediction and position our work within this context.

2.2.1 News Dissemination on Social Media

Social media is a very common channel for sharing news content. For example, 62% of American adults get news from social media (Reddit, Facebook, and Twitter are used most often); and among social media users, 80% often or sometimes click on links to news stories, 58% 'like' news stories, and 49% share or report news stories (Mitchell et al., 2016). These findings have also been confirmed for other countries. For example, BBC reports that social media is young people's main source of news[9], which makes social media crucial to reach that demographic. Hermida et al. (2012) report that 43% of Canadians who use social media receive news from a social media platform daily.

Twitter has been shown to be a major vehicle for news adoption and dissemination, both from the point of view of journalists, as well as computer scientists. Bruns and Burgess (2012) present a general overview of Twitter and its potential uses in news-related research, which include discussion of newsworthy events and 'gatewatching', i.e. identification and sharing of relevant material; and general commentary for current events. Kwak et al. (2010) show that unlike human social networks Twitter displays low reciprocity between users (i.e. asymmetrical follower/followee relationships) more characteristic of news outlets.

[9] http://www.bbc.com/news/uk-36528256 [Accessed 13th April 2018]

Furthermore, in their dataset covering 4,262 trending topics and 106 million tweets, the majority of trending topics (over 85%) were related to news. Twitter as a vehicle for news dissemination has been discussed widely in literature and much of the work has focused on analysing diffusion patterns for various hashtags and/or topics according to their temporal, topological, and content features (Lerman and Ghosh, 2010; Starbird and Palen, 2012). Most studies analyse only diffusion within one network (e.g. Twitter), however there is limited literature (e.g. Kim et al. (2012, 2013)) that analyses and predicts diffusion between networks (namely news, social networks, and blogs). In our work we take the latter approach and consider cross-domain popularity, i.e. popularity of news articles on social media.

Although Twitter is often used to disseminate news, Phillips (2012) report that in 2011 Facebook was responsible for referring twice as many people to four major news organisations as Twitter. Facebook driving more referrals to news websites than Twitter was also the finding of Olmstead et al. (2011). The significant impact that social media has on attracting readers to a news website means that news outlets should at least bear in mind what factors encourage readers to share news articles on social media (which is how they then act as referrals). A media blogger on *The Guardian* even suggests that "The key question for news organisations, tied to the goal of big traffic, is now 'what works best on Facebook?"'[10]. This makes our research crucial for news outlets, as we can provide insights into news article popularity on social media (and particularly how rewording headlines can lead to wider readership). Furthermore, our decision to use both Twitter and Facebook measures to model popularity of news articles allows us to explore the differences between the two social media websites (cf. Sections 3.2 and 6.5). In the following section we present an overview of the research on news article popularity prediction and position our work within this strand of research.

2.2.2 News Article Popularity Prediction

News article popularity prediction has been the subject of research for many years. There are several factors that can contribute to the relative popularity for this research task. Firstly, reading news online has become a daily habit for many Internet users. This leads to a large amount of data being generated. Secondly, some of that data is made openly available through Application Programming Interfaces (APIs) which facilitates large-scale data collection. Finally, this type of research has potentially far-reaching impact – both in terms of impact on online news readers (e.g. through better news recommender systems) and

[10] https://www.theguardian.com/media/media-blog/2015/may/17/facebook-news-digital-skills-automation [Accessed 13th April 2018]

impact on news outlets (e.g. through being able to better attract and engage with online readers).

2.2.3 Overview of Approaches

In Table 2.1 we present a selection of representative approaches to news article popularity prediction. We differentiate between these approaches in terms of:

the source of news articles; e.g. individual news outlets, news aggregators

the popularity measure; e.g. domain-internal, cross-domain; where **domain-internal** refers to using a popularity measure from the same source as the content (e.g. website comments for news articles), and **cross-domain** refers to using a popularity measure from outside of the source of content (e.g. tweets or Facebook shares for news articles)

the type of features; e.g. content, context; where **content** refers to any information contained within the content (e.g. text length), or that can be derived from the content (e.g. text sentiment), and **context** refers to information obtained from external sources (e.g. user or author characteristics, or popularity development patterns after publication)

data availability; e.g. pre-publication, post-publication; where **pre-publication** refers to models that make use of data which is available before the release of the content online, and **post-publication** refers to when the data needed for prediction only becomes available after publication (such as early popularity development)

Use of news sources. As shown in Table 2.1 there is considerable variety across the factors that we consider. Out of the ten reviewed approaches, six have used a news aggregator website. The most commonly used news aggregator was Digg (four instances) with single uses of Feedzilla and Yahoo News. The other approaches used articles from individual news outlets; either by using just one news outlet (two approaches), or several (also two approaches). We decided to collect data from two news outlets. This allows us to control the news source (thus reducing the number of confounding factors in our analysis), as well as compare the results for the two datasets (cf. Chapters 6 and 8), which can indicate differences in style or readership.

Use of popularity measures. In terms of popularity measures that were used, most approaches utilised a domain-internal measure of popularity, such as page views, number of comments under the news article, or the score on Digg. There were some cases of using

Table 2.1: Summary of news article popularity prediction research differentiating between: (*i*) sources for news articles; (*ii*) popularity measures; (*iii*) type of features; and (*iv*) data availability.

Authors	**News source**	**Popularity measure**	**Features**	**Time**
Jamali and Rangwala (2009)	Digg news aggregator	Digg score	**context**: comment popularity, comment length, user characteristics	post-publication
Lerman and Hogg (2010)	Digg news aggregator	votes on Digg	**context**: voting patterns	post-publication
Tsagkias et al. (2009)	seven news websites	comments on website	**content**: article length, some lexical features, number of entities; **context**: time of publication, presence of summary	pre-publication
Lerman and Hogg (2010)	Digg news aggregator	Digg score	**context**: voting patterns	post-publication
Bandari et al. (2012)	Feedzilla news aggregator	tweets	**content**: category, sentiment, prominence; **context**: news source	pre-publication
Ahmed et al. (2013)	Digg news aggregator	user votes	**context**: popularity patterns	post-publication
Hsieh et al. (2013)	*New York Times* articles	tweets	**context**: expert users, crowd wisdom	post-publication
Castillo et al. (2014)	*Al Jazeera* articles	pageviews	**context**: visitation patterns	post-publication
Tatar et al. (2014)	two news websites	comments on website	**context**: popularity distribution and lifespan	post-publication
Arapakis et al. (2014)	Yahoo News aggregator	tweets, pageviews	**content**: article length, genre, part-of-speech proportions, sentiment, # entities; **context:** time, news source	pre-publication

popularity measures external to the news source, namely popularity on social media. Two social media websites were used: Twitter and Facebook. Nearly all approaches tried to predict just one popularity measure with the exception of Arapakis et al. (2014) who used two – one domain-internal (pageviews) and one social media popularity measure (tweets). We use two external popularity measures: one from Twitter and one from Facebook. We combine the different types of popularity measures from the same domain (i.e. tweets and retweets on Twitter, and likes and shares on Facebook) to obtain an *overall measure of popularity* on a given social media website. This allows to target the whole social media network when writing headlines and not a specific sharing behaviour.

Use of content vs. context features. We categorised the features that were used for the task of news article popularity prediction as either content or context. All ten approaches used some form of context features. The majority (six approaches) used post-publication popularity patterns; that is to say, once an article is published how does the popularity of that article develop over time. Based on the early popularity patterns (e.g. within 30 minutes of publication), future popularity is extrapolated. The use of content features was less common (three approaches). Among content features that were used there were both linguistic aspects such as text length or proportions of various parts of speech, as well as extra-linguistic aspects like popularity of entities. In Appendix F we present an overview of the content features used in two state-of-the-art approaches which we later use as baselines in our prediction experiments (cf. Chapter 8)[11]. We add to these approaches by: (i) improving on existing features such as Prominence, (ii) adding new features (news values: Magnitude, Proximity, Surprise, Uniqueness; and style: Simplicity, Unambiguity, Punctuation), (iii) applying our methods to headlines only, and (iv) building a source-internal prediction model.

Data availability. As for the availability of the data used for prediction, the majority (seven out of ten) of approaches used data that is available only once the article is published, such as early popularity patterns. For example, Castillo et al. (2014) built a traffic prediction model for an online news outlet and achieved approximately R^2 = 0.7 in explained variance by using popularity patterns from the first 30 minutes after publication. However, using post-publication data is not always possible or appropriate, as content authors might benefit from an expectation of popularity *before* they publish their content. This scenario when post-publication data is not used for prediction is analogous to the new item 'cold-

[11] We note that Arapakis et al. (2014) was recently updated and published as Arapakis et al. (2017). We compare our approach against the earlier version, but there were no significant changes to their approach which would impact the comparison with our model. The new version adds new popularity measures from Facebook (which we also consider) and reports feature correlations (which we do for a larger number of text-derived features, as well as investigate correlations with perceived popularity)

start' problem in recommender systems, when a new content item is entered into the recommender system, but no usage data is available for it yet (Son, 2016). Such 'cold-start' systems need to rely on other types of data in order to make a prediction or a recommendation. In the approaches that we review there are three that used features which were all available before publication – mostly content features derived directly from article text, or from metadata about the article such as what time of day the article will be published. In these model the explained variance is considerably lower (e.g. R^2 = 0.43 in Bandari et al. (2012)). In order to account for cases when post-publication data is not available or cannot be used, it is crucial to establish a reliable method for pre-publication, or 'cold-start' prediction. In this thesis we investigate the predictive power of news values and style features derived from headlines (cf. Chapter 8) which can be used for a 'cold-start' problem.

In our project, we conduct a *source-internal* investigation of headline-derived *content features* (available *pre-publication*) using *cross-domain* popularity measures from Twitter and Facebook. This allows us to test our main hypothesis that the way we formulate headlines influences the popularity of news articles on social media.

We collect data from **two news outlets** (*The Guardian* and *New York Times*) and conduct our analysis on each source separately. This ensures that news source popularity effects observed in Bandari et al. (2012) and Arapakis et al. (2014) are not present. Furthermore, we can test the generalisability of our feature extraction methods (cf. Chapters 4 and 5), as well as compare and contrast the results of applying our methods on two different corpora (cf. Chapter 6).

The popularity measures we use are obtained from two social media websites: Twitter and Facebook. We chose **cross-domain popularity measures**, because news outlets are increasingly interacting with social media (e.g. providing social media sharing options) and need a way of targeting social media users specifically (Olmstead et al., 2011). Cross-domain popularity measures have also been used less commonly in research on news article popularity prediction, despite the clear impact that social media websites have on news dissemination (cf. Section 2.2.1).

The features we use are **derived from headline text**, with a limited number of article metadata features used in the prediction models. This is motivated by a relatively minor role that content features play in the task of news article popularity prediction (they were used in three out of ten approaches presented here), and the fact that only a limited number of text-derived features have been considered for this task (cf. Appendix F). Furthermore, the features that have been considered were only applied to all sections of the article (title and main body). However, we argue in the following section (cf. Section 2.3) that headlines play a special role in news discourse and should be considered separately. As far as we are aware, headlines on their own have not been used for the task of news article popularity prediction. In this thesis we explore a wide variety of headline-derived features (news values inspired by journalism studies in Chapter 4, and style features from natural language processing literature in Chapter 5) and their impact on social media popularity of news articles.

Related to this is our decision to use only data that is available **pre-publication**. The reason for that is that this task formulation remains relatively unexplored in the news article popularity prediction field (with some exceptions like Tsagkias et al. (2009); Bandari et al. (2012); Arapakis et al. (2014)). Furthermore, models using only pre-publication data would be a useful tool when generating news or news-like content, so that social media response can be predicted without actually releasing potentially sensitive information (such as would be found in news scoops, or marketing campaigns).

Compared to previous work we contribute by: (i) reporting a more robust investigation of content features by conducting a source-internal analysis, (ii) considering a wider range of content features which can be used in a 'cold-start' problem, (iii) investigating in detail the impact of individual features on social media popularity, and (iv) focusing on headline text.

2.3 The Importance of Headlines

The previous section presented an overview of approaches to news article popularity prediction. All approaches which used text-derived features used the whole article text. In this section we present evidence that headlines play a special role in news discourse and should be studied as a separate phenomenon. In this thesis we ask whether features derived only from *headlines* are sufficient for modelling popularity of news articles on social media.

2.3.1 Headlines in the Digital Age

There is a tremendous amount of digital content available nowadays (cf. Section 2.1.1). A significant proportion of that content features some form of title or a headline, and they play a vital role in attracting audiences' attention to online artefacts.

Literature on information retrieval (IR) provides some insights into the role of headlines. Information retrieval systems pay particular attention to the title of a document, as well as to the opening sentence. Titles and opening sentences are known to play an important rhetorical function in text (cf. literature on summarisation e.g. Radev et al. (2002)). When it comes to retrieving and/or summarising online news stories, one of the key features to be exploited are headlines (Tran et al., 2015). Headlines are written so that they are memorable to the readers (Perfetti et al., 1987), and provide a high-level overview of the story without demanding much of the reader's time or attention (Dor, 2003). IR systems which process online news stories are likely to rely on headlines and opening sentences, since the writing style and general approach to covering news might vary across geographical regions and individual news outlets, but their position in news text is always the most prominent and important for conveying the message (Radev et al., 2005).

Headlines were found to be one of main visual entry points to online news content (Leckner, 2012). This is intensified on social media platforms, where in cases of indirect engagement (e.g. with retweeted news articles) headlines are often the only visible part of the main content, and hence play a key role in attracting attention. Liu (2005) found that compared to print media, digital readers spend more time browsing and scanning, keyword spotting, and on selective and non-linear reading. Various studies conducted by Chartbeat, an analytics company that focuses on online attention metrics, found that 38% of users leave a website immediately after accessing it[12], and that an average reader will spend only 15 seconds on a website[13]. An American Press Institute study found that roughly six in 10 people acknowledge that they are 'headline-gazers' checking only the headline and not reading the full article[14].

Eye-tracking studies have also confirmed the importance of headlines. They have shown that in online browsing and searching contexts readers attend to what is prominently displayed to them. For example, when using Google search, users' attention is mostly focused on the top texts in a ranked list, and furthermore, within a given text it is focused on the first sentences (Pan et al., 2007; Lorigo et al., 2008). Eye-tracking studies focused specifically on online news reading have shown that many people are 'entry-point readers' who attend to headlines in order to ascertain the overview of an article, but who exhibit minimal reading activities (Holmqvist et al., 2003; Holsanova et al., 2006).

Headlines are also gaining ground in the natural language processing community as a text type to be studied separately from full articles. This includes work on headline generation (Kourogi et al., 2015; Gatti et al., 2016) and keyword selection for popularising content (Szymanski et al., 2016). Chesney et al. (2017) recently looked at the incongruent nature of headlines (i.e. headlines presenting a biased view of the main content).

In this thesis we investigate in detail how certain headline characteristics (news values and linguistic style) influence the social media popularity of news articles. The research presented in this section which points to the commonness of 'headline-gazing' and 'entry-

[12] http://slate.me/1cJ7b5C [Accessed 13th April 2018]

[13] http://yhoo.it/2cEQMVC [Accessed 13th April 2018]

[14] http://bit.ly/21LwfS5 [Accessed 13th April 2018]

point reading' makes our analysis especially timely, since we provide insights about how to formulate a headline to make it more attractive to social media users.

2.3.2 Challenges for Automatic Processing

From a natural language processing perspective headlines pose a computational challenge due to certain linguistic aspects like unusual use of tenses (Chovanec, 2014) and deliberate ambiguity (Brône and Coulson, 2010). Moreover, headlines are typically short, which severely limits the amount of context that many NLP tools rely on. While feature engineering for headlines is less studied, there are research efforts in the NLP community that specifically address different types of short texts. Tweets have attracted considerable attention, leading to the development of some Twitter-specific tools (e.g. TweetNLP[15]). Tan et al. (2014) is an example of extracting linguistic features from tweets. Another example of a text type that is closely related to headlines are online content titles, e.g. Reddit titles (Lakkaraju et al., 2013). Overall, various text characteristics were used for short texts – either content-specific (e.g. including explicit requests for sharing in tweets (Tan et al., 2014)), or more general, like ratios for various parts of speech, sentiment, similarity to a language model. The latter ones are more widely applicable, but need to be adjusted to work with headlines. For example, since headlines offer limited context, sentiment analysis carried out on word-level is more appropriate (cf. Tan et al. (2014), Gatti et al. (2016), Szymanski et al. (2016)). Clickbaiting ("headlines which are sensationalised, turn out to be adverts or are simply misleading")[16] applies to headlines in many online news outlets (e.g. Buzzfeed is the commonly named example) and has a linguistic realisation (e.g. forward-referring deixis such as "*This* news will surprise you" (Blom and Hansen, 2015), or 'listicles' which begin with a cardinal number[17]). Some of the clickbait-related linguistic features (personal and possessive pronouns, sentimental words, quote marks, question marks) have been found to correlate with higher/lower click-through rates in a sample of Dutch newspaper article headlines (Kuiken et al., 2017). As our datasets consist of broadsheet newspapers, which do not tend to use clickbait, we do not target clickbait explicitly in the feature engineering, however some features we do consider (e.g. Sentiment) are common between broadsheet and clickbait headlines.

In this thesis we develop our own methods, as well as make use of a number of state-of-the-art tools to process the headlines corpora (cf. Chapters 4 and 5). In many cases (e.g. Prominence, Sentiment) our approach is influenced by methods applied to short texts other

[15] http://www.cs.cmu.edu/ ark/TweetNLP/ [Accessed 13th April 2018]

[16] http://www.bbc.co.uk/news/uk-wales-34213693 [Accessed 13th April 2018]

[17] https://www.wsj.com/articles/buzzfeed-nails-the-listicle-what-happens-next-1422556723 [Accessed 13th April 2018]

than headlines, which we cover in the following section.

2.4 The Effect of Wording in Short Texts on Online Popularity

When looking at the various approaches to news article popularity prediction (cf. Section 2.2.2) we noted a range of features which were used and categorised them as content or context features. Although content features are used less often, they are crucial for predicting popularity of online artefacts using only data that is available before publication, in our case using headline text to predict the popularity of news articles on social media. In journalism studies, the linguistic style of a news article has been seen as an important part of the news production process. For example, the role of "different choices of words and grammatical phrasings" has been mentioned by Fowler (1991, p.66). In order to consider linguistic style of headlines for modelling social media popularity, we need to decide which headline features should be investigated. Relevant to feature engineering using textual content are the research efforts which focus on the effect of wording or phrasing on popularity. This earlier research informs our choice of features, in particular the linguistic style features. In this section we provide an overview of that literature, focusing on short texts, since the features used in these approaches are relevant to our choice of features for modelling news article popularity using headlines.

There have been a number of research efforts to establish the effect of wording on different types of texts. Some examples use longer pieces of text. For example, Ashok et al. (2013) used a number of style features to predict the success of novels. When comparing the style of more successful with less successful novels they found that some unigrams and part-of-speech distributions are discriminative. Interestingly they found that for discriminative unigrams high sentiment was more common for less successful novels and that readability and literary success had a negative association. Guerini et al. (2012) and Louis and Nenkova (2013) looked at science writing. Guerini et al. (2012) found that linguistic style and readability of abstracts affected the popularity (measured in citations and bookmarks) of scientific articles. Louis and Nenkova (2013) explored style and wording of popular science articles. The factors that they found to positively influence science article popularity were using visual words (i.e. words evoking an image, such as *grass*, *blue*, *diamond*, *dots*, *carousel*) at the beginning and end of the article, creative language, and affective content. The challenge of using textual content of longer documents (especially novels) is finding features which accurately summarise the document's style and content. Discourse also comes into play (e.g. Pitler and Nenkova (2008) used discourse relations to predict text quality), but is challenging to reliably implement and is usually not present in short texts like headlines. Some of the general types of features successfully used in these approaches can be applied to headlines, for example: proportions of parts of speech (our model: Nouns, Verbs, Adverbs in Chapter 5), affective content (Sentiment in Chapter 4), creative language (Surprise in Chapter 4).

Because headlines tend to be short, we are particularly interested in research efforts which used content of short texts to predict their popularity. This allows us to identify features which can be applied on headlines and used to model social media popularity of news articles.

One type of short text that has garnered considerable research attention is tweets. While many tweet popularity approaches make use of social graph or temporal features (which we call context features), there is a small number of approaches which focus on features derived from the textual content of tweets. One approach which used only content features was by Tan et al. (2014). Their goal was to determine what effect tweet wording has on popularity. Due to the nature of Twitter usage where users slightly edit and repost messages, they were able to find similarly worded tweet pairs, thus controlling for topic and author effects. They found a number of factors that were positively associated with popularity, for example: explicit requests for sharing, tweet length, conforming to expectations (based on a language model), words with positive or negative connotations, and readability. Because of the ability to control for topic and author, these results are some of the clearest indicators of the effect of tweet wording on its popularity. However, most types of online content do not present the same opportunity to control for multiple factors which leads to some noise being introduced. Since the features in Tan et al. (2014) were shown to have a significant correlation with Twitter popularity (which is one of the social media popularity metrics we use for modelling), we reimplement some of their features in our model, e.g. length (our model: Brevity in Chapter 5), conforming to expectations (Simplicity in Chapter 5), and connotations (Sentiment in Chapter 4).

Other types of short texts have been considered as well. Danescu-Niculescu-Mizil et al. (2012) investigated factors which affect memorability of film quotes and found that lexical distinctiveness but syntactic similarity (based on a language model) had a positive effect. Lakkaraju et al. (2013) looked at pairs of Reddit image titles (i.e. the same image submitted with two different titles). They found that titles very similar to past submissions perform badly and that using nouns and adjectives has a greater impact on popularity than verbs and adverbs. Reis et al. (2015) looked at the effect of sentiment in headlines on news article popularity. They found that extreme sentiment scores had the largest mean popularity, suggesting that both strongly negative and strongly positive headlines tend to attract more readers. Both Uniqueness and Sentiment features are included in our model (cf. Chapter 4).

Despite short texts (especially tweets) having attracted quite significant research attention, as far as we are aware there have been no attempts until now to model news article popularity using only headlines. In this thesis we investigate news values and linguistic style features for this task. The implementation for some of them is informed by the research presented in this chapter. We build on this research by exploring in depth a wide range of features, which can be applied to other types of short texts. We also implement a group of features which offer a journalistic perspective – news values. They are presented in the next section.

2.5 News Values and the Journalistic Perspective

A variety of text-derived features have been used for modelling the popularity of online content (cf. Section 2.4) and online news content in particular (cf. Section 2.2.2). In this thesis we propose a new type of features – news values – which can be extracted from text using computational methods. By using news values we add a *journalistic perspective* to the task of modelling social media popularity of news articles, which has not been formalised in a computational approach for this task before. O'Neill and Harcup (2009, p.172) argued that news values can be of use not only to journalists, but also to "PR professionals, critics of mainstream media, marginalised groups wanting to publicise their message, and citizens". This makes them a valuable contribution when modelling the popularity of online news articles, as well as online content in general. In this section we present an overview of literature on news values including a summary of selected news values taxonomies, which informs our choice of news values to implement. We also present some examples of using news values in research tasks and how these can be enriched by having a computational method for extracting news values from headlines.

2.5.1 Definitions and Taxonomies

News values originated in the journalism studies field with the work by Galtung and Ruge (1965). They analysed news articles on international crises in the Norwegian press and hypothesised that if an event can be characterised by one or more of the twelve factors they propose, this event will become news. The twelve factors they proposed are:

Frequency: news events which follow the frequency/cycle of news media

Threshold: news events with certain threshold of size or intensity

Unambiguity: news events written in a way that leaves no ambiguity

Meaningfulness: news events which are relevant to the audience

Consonance: news events which are predictable or follow a certain pattern
Unexpectedness: news events which are surprising or unpredictable
Continuity: news events which follow on from earlier reports
Composition: news events which are used to fill and balance the newspaper
Reference to elite nations: news events which refer to prominent nations
Reference to elite people: news events which refer to prominent people
Reference to persons: personalising news events using everyday people
Reference to something negative: news events which refer to something negative

The news values taxonomies have not remained static. For example, Harcup and O'Neill (2001) analysed 1276 page leads from three British newspapers and found that some of the features proposed by Galtung and Ruge still characterise a large number of news articles. The authors also included some new features, such as Entertainment to reflect the changes in the news discourse since the article by Galtung and Ruge. They also repeated their analysis of Galtung and Ruge more recently (Harcup and O'Neill, 2016) and investigated how the prominence of various news values might have changed over the years. They closed with an updated taxonomy of news values.

A number of news values taxonomies have been proposed over the years: Bell (1991), Harcup and O'Neill (2001), Johnson-Cartee (2005), Bednarek and Caple (2012), and Harcup and O'Neill (2016). Although there are some differences in the number of concepts, granularity and definitions of news values, there is in fact a considerable overlap between all these taxonomies (Caple and Bednarek, 2013). We present a summary of selected news values taxonomies in Table 2.2.

In our work we operationalise news values (cf. Chapter 4), which: (i) occur in more than one taxonomy, (ii) are realised explicitly through headline language, and (iii) can be clearly defined. We also follow (Bednarek and Caple, 2012, p.41) in distinguishing between news stories (realisation of news events in text), news writing objectives (the style of news story text), and selection factors (news agenda and news cycle). Following that categorisation, news values only apply to news stories. Several factors which appear in news values taxonomies we presented in Table 2.2 do not refer to news stories and thus we do not operationalise them as news values. For example, Unambiguity in Galtung and Ruge (1965) and Brevity in Johnson-Cartee (2005) are news writing objectives, which we implement as linguistic style features (cf. Chapter 5). Composition in Galtung and Ruge (1965), and newspaper agenda and exclusivity in Harcup and O'Neill (2016) refer to selection factors, which we do not consider in this work. Some news values cannot be clearly defined, which excludes them from operationalisation; e.g. Shareability – meaning

Table 2.2: Overview of selected news values taxonomies.

Galtung and Ruge (1965)	**Johnson-Cartee (2005)**	**Bednarek and Caple (2012)**	**Harcup and O'Neill (2016)**	**Our work**
Frequency				
Threshold	Size	Impact, Superlativeness	Magnitude	Magnitude
Unambiguity				
Meaningfulness	Social Impact, Proximity	Proximity	Relevance	Proximity
Consonance	Familiarity	Consonance		
Unexpectedness	Novelty	Novelty	Surprise	Surprise
Continuity			Follow-up	Uniqueness
Composition				
Ref. to elite nations		Prominence	Power elite	Prominence
Ref. to elite people		Prominence	Power elite	Prominence
Ref. to persons		Personalisation		
Ref. to something negative	Negativity	Negativity	Bad news	Sentiment
			Good news	Sentiment
			Celebrity	Prominence
			Entertainment	
			Newspaper agenda	
			Exclusivity	
	Conflict		Conflict	Sentiment
	Visual attractiveness		Audio-visuals	
			Shareability	
	Drama		Drama	Sentiment
	Timeliness	Timeliness		
	Action			
	Brevity			

shareability on social media (Harcup and O'Neill, 2016). Finally, Johnson-Cartee (2005) and Harcup and O'Neill (2016) propose including non-text media as a news values, however these fall outside of the scope of this thesis.

Within the NLP field, Arapakis et al. (2016) came up with a similar taxonomy and proposed a list of 14 news article quality aspects. Although some of them could be categorised as either news values (e.g. Novelty, Sentimentality) or news writing objectives (e.g. Conciseness, Formality), the authors did not make an explicit link with the journalism studies literature. In Chapter 4 we propose implementations for news values which have linguistic indicators in headline text following a review of news values taxonomies, and in Chapter 5 we implement a number of linguistic style features which also include news writing objectives.

Galtung and Ruge (1965) also hypothesised that the effect of news values is cumulative, that is to say, the more news values an event exhibits, the more newsworthy it is. We investigate this in Chapter 6 where we correlate feature values (including news values) with social media popularity metrics. Although the focus of this work is news values realised through explicit linguistic indicators, some news values require looking beyond the text. Some of the news values which are presented in Table 2.2 depend on the reader's location (e.g. Meaningfulness/Proximity, Reference to elite people). We take this into account in Chapter 9 where we build a country-specific popularity prediction model.

In this section we presented an overview of news values and their taxonomies. We differentiated between news values (which refer to news stories) and news writing objectives (which refer to linguistic style). Our criteria for selecting news values for the task of modelling social media popularity of news articles using headlines led to the selection of six news values which we summarise in section 2.5.2 and describe their operationalisation in Chapter 4. We summarise news writing objectives in Section 2.5.3 and their implementation as linguistic style features in Chapter 5.

2.5.2 Overview of Selected News Values

Prominence. Referring to prominent entities is one of the key news values and appears in almost all news values taxonomies. Galtung and Ruge (1965) listed Reference to elite nations and Reference to elite people as part of this news value. Harcup and O'Neill (2001) revised these news values as Power Elite and updated their taxonomy to include reference to celebrities. Prominence can be interpreted as eliteness (Reference to elite people and Reference to elite organisations in (Galtung and Ruge, 1965, p.68); Power elite in Harcup and O'Neill (2016)), or recognisability ("objects of general identification"; Galtung and Ruge (1965, p.68)).

Sentiment. This refers to sentiment-charged events (Johnson-Cartee, 2005) and using sentiment-charged language (Bednarek and Caple, 2012). In particular, negative sentiment has been considered a common feature of news discourse (Johnson-Cartee, 2005; Bednarek and Caple, 2012; Harcup and O'Neill, 2016). Gans (1979, p.52) suggested that there is a journalistic bias towards extreme behaviours or concepts expressed through 'pejorative adjectives', which can indirectly express sentiment. Sentiment in news has also been studied from the computer science perspective, e.g. the significant impact of sentiment and emotionality on virality (Berger and Milkman, 2012). Reis et al. (2015) who looked at the effect of sentiment on headline popularity found that headlines with extremely positive and extremely negative sentiment tend to be more popular for online news.

Magnitude. The size (Johnson-Cartee, 2005, p.128), or magnitude (Harcup and O'Neill, 2001) of an event is considered to influence newsworthiness. In terms of linguistic expressions of Magnitude Cotter (2010, p.161) pointed out the use of superlative adjectives and Bednarek and Caple (2012, p.47) the use of intensified vocabulary in news discourse.

Proximity. Proximity has been linked to relevance – the justification being that news events which are 'close' to the reader are more newsworthy. In journalism studies literature Proximity commonly refers to geographic proximity (Johnson-Cartee, 2005, p.128). However, cultural proximity (Galtung and Ruge, 1965; Gans, 1979), which takes into account possible relevance of news items in the context of the wider cultural context, is also considered.

Surprise. Events which involve "surprise and/or contrast" (Harcup and O'Neill, 2001) make news. This news values has also been termed as Unexpectedness (Galtung and Ruge, 1965) or Novelty (Johnson-Cartee, 2005; Bednarek and Caple, 2012). Usually this news value is defined as the news event being surprising or unexpected (Johnson-Cartee, 2005, p.128). For example: "Four-metre shark spotted off WA coast placed on fisheries kill list", "Denver Post hires Whoopi Goldberg to write for marijuana blog", or "Britain's first cloned dog born after £60,000 test-tube procedure". These examples require world knowledge to identify Surprise. We call this type of Surprise *implicit surprise*. In this thesis we look at *explicit surprise*, i.e. surprising phrasing which provides an explicit linguistic indication of Surprise. For example: "Spanish town hires its own *pet (poo) detective*", and "Beekeeper creates *coat of living bees* - in pictures".

Uniqueness. News has to be new – "any new comment or circumstance [...] adds to the debate" (Conley and Lamble, 2006). Although Continuity (Galtung and Ruge, 1965) and Follow-up (Harcup and O'Neill, 2016) suggest that a news story continuing an earlier storyline is newsworthy, an analysis of several storylines in our data revealed the opposite (cf. Section 4.1.6). We investigate which headlines (unique or similar) are more likely to be shared on social media (cf. Chapter 6).

2.5.3 Overview of Selected News Writing Objectives

While various taxonomies exist for news values (cf. Table 2.2), there is less clarity about the role of news writing objectives. We use two sources to gather information about news writing objectives: (i) a list of news writing objectives in news values literature provided by Caple and Bednarek (2013), and (ii) news outlet guidance for writing headlines obtained from *The Guardian* style guide[18] and *Yahoo!* style guide (Barr, 2010). We categorised them into the following seven groups.

Brevity. Traditionally space is limited in newspapers (Bell, 1995). This has led to the need for brevity when writing headlines (Dor, 2003; Cotter, 2010).

Simplicity. Simplicity, or 'ease of comprehension', has been suggested as one of the objectives of news writing (Cotter, 2010). Bednarek and Caple (2012, ch.4) state that this relates to various linguistic aspects of news writing, such as syntax or vocabulary. *Yahoo!* style guide advises clarity over cleverness when writing headlines.

Unambiguity. The news writing objective of Unambiguity is highly related to Simplicity, as simpler writing will tend to be less ambiguous. Unambiguity has been considered important for news writing since Galtung and Ruge (1965). Avoiding ambiguity, especially in relation to successive nouns (which ties in with the section on Nouns below), is recommended in *The Guardian* style guide. *Yahoo!* style guide says: "Inaccurate or misleading headlines are worse [than not successful headlines]".

Punctuation. The usage of punctuation marks is covered by the style guides on headlines. *The Guardian* style guide states that certain punctuation marks, namely quote marks, question marks, and exclamation marks, are to be avoided: "Exclamation marks – look, I've written something funny! – should never be used. Question marks are also to be avoided, as are quotation marks, unless essential to signify a quote or for legal reasons". *Yahoo!* style guide offers similar guidance: "If you want to include question marks and exclamation points, be stingy with them".

Nouns. *The Guardian* style guide discourages using too many successive nouns (so-called 'headlinese', e.g. "New York assault weapons ban") in order to avoid ambiguity (for example, "Landmine claims dog UK arms firm"). Interestingly, *Yahoo!* style guide advises the use of proper nouns, which can help with search engine optimisation for the headline.

[18] https://www.theguardian.com/guardian-observer-style-guide-h [Accessed 13th April 2018]

Verbs. Using verbs is encouraged in headlines in both *The Guardian* and *Yahoo!* style guide.

Adverbs. Adverbs, especially adverbs of manner, are frequently used in headlines (Bednarek and Caple, 2012).

We implement the news writing objectives presented in this section as linguistic style features in Chapter 5.

2.5.4 Use of News Values in Research

News values have been widely used in journalism studies, however researchers still mainly rely on manual annotation. For example, news values were used by Bednarek and Caple (2014) to analyse news discourse, while Kepplinger and Ehmig (2006) used them to predict the newsworthiness of news articles. Since news values need to be annotated manually, large-scale analyses of news articles in the journalism studies field have focused on more general aspects of news articles that are readily available through article metadata (such as topics, e.g. Bastos (2015)).

There have been some limited attempts at using computational methods to enable large-scale annotation of news values from text, however these can be described at most as semi-automatic, since they require significant manual effort. The approach by Potts et al. (2015) is an example of such a hybrid approach. First, they used some techniques from corpus linguistics (part-of-speech and semantic tagging, and collocations and lemma frequencies based on those) to analyse a large news corpus about hurricane Katrina. Then they obtained lists of most frequent tags and collocations. Following that they *manually* chose words which indicate news values. While it is a step towards large-scale studies using news values, this approach approach does not generalise well, as keywords which indicate news values are specific to the topics covered in the corpus.

More recently di Buono et al. (2017) attempted to predict news values from headline text using word embeddings and emotion features. The classification results varied depending on the news value from an F1 score of 0.43 to 0.85. Although this was a fully automatic approach to deriving news values from headlines, it required a significant amount of manual annotation to obtain the classification labels.

We contribute to this research direction by developing fully automatic computational methods for extracting news values from headline text (cf. Chapter 4). Such computational methods can be used for large-scale investigations into the prevalence of news values across different topics, genres, or news outlets (cf. Section 10.4). Unlike (di Buono et al., 2017) who used word embeddings and emotion labels, our approach offers *interpretable* results. For example, knowing that a particular feature is positively correlated with social media popularity, the headline author can edit their headline in such a way that maximises that feature.

2.6 Summary

The subject of this thesis is modelling social media popularity (global, as well as country-specific) of individual news articles using news values and linguistic style features derived from headline text.

In this chapter we presented an overview of the research on predicting popularity of online content, focusing on the task of **news article popularity prediction**. We argued for considering **headlines** separately for this task based on the research findings related to the special role and function of headlines (e.g. readers are 'headline-gazers'). To address our requirement for the use of **content features** which allows to make a prediction prior to content publication, we summarised the literature on the effect of **short text wording** on online popularity. This inspired the implementation in our model of the features which were found to have significant impact on popularity. Finally, we introduced **news values** which provide previously not considered journalistic perspective to the task of online news article popularity prediction. We also motivated our choice of news values to operationalise in our models and provided their summaries.

We enhance prior work on news article popularity prediction by: (i) using only headline text; (ii) operationalising and evaluating news values for this task; (iii) adding a detailed investigation of news values and style features' impact on social media popularity and on perceived popularity; and (iv) building country-specific as well as global prediction models.

In order to develop and evaluate our proposed methods for modelling the social media popularity of news articles using headlines we need corpora which consist of news article headlines which are associated with measures of social media popularity. We present the data collection and preprocessing methods in the next chapter.

Chapter 3

Data Collection and Preprocessing

Our goal is to model social media popularity of news articles using headlines. To achieve this we follow an experimental methodology whereby we develop feature engineering methods for news values and linguistic style and apply them on corpora consisting of news article headlines. The social media popularity metrics associated with these news articles are used to investigate the individual and combined impact of these features. In this chapter we give details of the data collection process for headlines and the associated social media popularity metrics, as well as outline the preprocessing that was carried out on the headlines corpora. The datasets we created are made publicly available (cf. Appendix A).

3.1 Headlines Corpora

To create the headlines corpora we used data from two news major global outlets – *The Guardian* and *New York Times*. As of July 2017, *The Guardian* had an online readership of approximately 1.2 million[1], while *New York Times* reported 2.2 million digital-only subscriptions[2]. As major news outlets, *The Guardian* and *New York Times* provide us with a wide coverage of various topics and genres, which allows for a good exploration of the impact of news values and linguistic style. This also provides corpora from two established broadsheet news outlets, allowing a comparison of similar content from different locations.

[1] http://www.newsworks.org.uk/The-Guardian [Accessed 13th April 2018]

[2] https://www.nytimes.com/2017/05/03/business/new-york-times-co-q1-earnings.html [Accessed 13th April 2018]

3.1.1 Collection

We downloaded all headlines using publicly available APIs: *The Guardian* Content API[3] and *New York Times* Article Search API[4]. We also collected some of the metadata associated with the articles. For *The Guardian* we collected: article identifier, category, tags. For *New York Times* we collected: article identifier, section, and genre.

To ensure a robust evaluation of the prediction models (cf. Chapter 8), we collected headlines data during two distinct periods for each news outlets, thus creating training and test sets separated by at least a month. The collection was carried out throughout 2014. First for *The Guardian* in April (*The Guardian* training set) and July (*The Guardian* test set), then for *New York Times* in October (*New York Times* training set), and December (*New York Times* test set). In total we collected 25,786 articles from *The Guardian* and 10,085 articles from *New York Times*. The sizes of the training and test sets for each news outlet are comparable (we provide a more detailed overview of the collected corpora in Section 3.1.3). A summary of the collection process is given in Table 3.1.

Table 3.1: Summary of the data collection process for headlines corpora.

		The Guardian	*New York Times*
Collection method		API	API
Collection period	Training	April 2014	October 2014
	Test	July 2014	December 2014
Number of articles	Training	11,980	5,074
	Test	13,806	5,011

Any duplicated headlines (i.e. with the same URL) or articles which were removed (indicated as "Removed:" in headline) were removed from the datasets.

3.1.2 Preprocessing

Part-of-speech tagging and parsing. As a first step all headlines were part-of-speech tagged using the Stanford Part-of-Speech Tagger (Toutanova et al., 2003) and parsed using the Stanford Parser (Klein and Manning, 2003). Both tools were developed and trained on newswire datasets. The POS-Tagger achieved 97.24% token accuracy and the Parser achieved 86.32% F1 score.

Wikification. We decided to use wikification (a method of entity linking which connects keywords in text to the relevant Wikipedia page; e.g. Mihalcea and Csomai (2007)) to

[3] http://www.theguardian.com/open-platform [Accessed 13th April 2018]

[4] http://developer.nytimes.com/docs [Accessed 13th April 2018]

identify entities in the text. This allows us to explore a wider range of entities (e.g. concepts, titles) beyond the Person, Location, Organisation entity set which is commonly used in standard named entity recognisers. By linking entities to Wikipedia pages we can also access Wikidata, the knowledge graph behind Wikipedia, which facilitates our experiments on country-specific popularity prediction in Chapter 9. Headlines were wikified using the TagMe API[5]. It is a tool meant for short texts, making it suitable for headlines. In an evaluation of seven entity linking systems by Cornolti et al. (2013) TagMe achieved the highest F1 measure for three newswire datasets (between F1 = 50.7 to F1 = 58.3 depending on dataset). It also achieved the highest F1 scores when considering mention matching (i.e. recognising entity mentions in text; F1 = 74.6) and entity matching (i.e. linking text match to Wikipedia page; F1 = 65.6). The TagMe output for a headline returns a set of entities (corresponding to Wikipedia pages) and Wikipedia categories for those pages.

Notation. The following notation is used throughout this thesis. One example of a preprocessed headline with the notation is presented in Headline 3.1. Further examples are presented in Appendix B.

- H refers to the set of tokens obtained from the part-of-speech tagger from the headline.
- C refers to the set of content words in the headline. We define a content word as a noun, verb, adjective, or adverb.
- E refers to the set of entities in the headline as identified by TagMe.

Headline 3.1. "Emma Watson's makeup tweets highlight the commodification of beauty"

$H =$ { *Emma*, *Watson*, *'s*, *makeup*, *tweets*, *highlight*, *the*, *commodification*, *of*, *beauty* }

$C =$ { *makeup*, *tweets*, *highlight*, *commodification*, *beauty* }

$E =$ { EMMA WATSON, COMMODIFICATION }

3.1.3 Overview of the Headlines Corpora

Headline length. Table 3.2 presents an overview of the number of words and the number of TagMe entities in the headlines corpora. Apart from the maximum value for the number of words in *The Guardian* dataset, there are no differences between training and test

[5] https://tagme.d4science.org/tagme/ [Accessed 13th April 2018]

datasets. This is a good indication that despite the different data collection periods, the data are comparable.

For both *The Guardian* and *New York Times* the minimum number of words is one (cf. Headline 3.2). However, *The Guardian* headlines are on average longer – both the median and maximum values are higher in *The Guardian* corpus than *New York Times* (cf. Headline 3.3 for the longest headline in the dataset).

Headline 3.2. “Chatterbox”

Headline 3.3. “Jeeves & Wooster in ‘Perfect Nonsense’: Join Robert Webb and Mark Heap for an exclusive post show Q&A on 6 May and get top price tickets for just £39.50”

The statistics for the number of TagMe entities are quite similar for the two corpora. The minimum and median numbers of TagMe entities are the same for both datasets (0 and 1 respectively). Once again *The Guardian* has a higher maximum value than *New York Times*, but this might be because the headlines also tend to be longer. Further examples of preprocessed headlines are included in Appendix B.

Table 3.2: Overview of the number of words and TagMe entities in *The Guardian* and *New York Times* corpora. Minimum (Min.), median (Med.), and maximum (Max.) are reported.

		The Guardian		*New York Times*	
		Training	Test	Training	Test
Number of words (H)	Min.	1	1	1	1
	Med.	10	10	8	8
	Max.	29	27	18	18
Number of TagMe entities (E)	Min.	0	0	0	0
	Med.	1	1	1	1
	Max.	8	8	4	4

Categories. When we think of news, what usually comes to mind are the headlines about events. They can be categorised as ‘hard news’. However, news outlets do not just report on events. Other categories like opinion, reviews, interviews, and even recipes are common. The prevalence and importance of news values and linguistic style features might differ between categories, which is why we take them into account in our analysis. In this section we present an overview of categories in the two headlines corpora.

One of the differences between *The Guardian* and *New York Times* corpora is how they tag news articles. In *The Guardian* API a category is provided, which can refer to a topic (e.g. *Technology*, *Society*) or genre (e.g. *Review*); sometimes even within the same category (e.g. *World news*). There are 160 distinct categories in *The Guardian* dataset we collected. In *New York Times* API two tags are supplied with each news article: one for topic (what they call *Section*) and one for genre. For example, within the same topical section *Arts* there are articles indicated as being of the genre *News* and *Review*. There are 27 distinct genres in the *New York Times* dataset. This difference in granularity of categories might have implications for the prediction models in Chapter 8 which include metadata such as category.

By using metadata supplied through the APIs we can obtain a subset of articles belonging to the *News* genre (as opposed to *Opinion*, *Review*, or *Recipe*). This news subset will be later used to investigate the impact of news values and style features on social media popularity in Chapter 6. We obtain the news subset for *New York Times* by selecting only articles whose genre is *News*. For *New York Times* training set this resulted in 3843 articles (out of 5074). For *The Guardian*, we obtain the news subset by selecting only articles which have the tag *News*, or a tag containing the token *news*. This resulted in identifying 3161 articles in *The Guardian* training set (out of 11980) as the news subset.

Table 3.3 presents an overview of the prevalence of the twenty most frequent categories in *The Guardian* and *New York Times* corpora. While news is the top item in both datasets, for *New York Times* corpus it is much clearer that news is the prevalent genre (63% of the dataset). The top three categories in *The Guardian* – *World news*, *Sport*, and *Football* – mostly contain news articles as well, however some of the sport articles in those categories belong to other genres such as opinion or interview. There is also an apparent overlap in coverage between the items in both corpora. For example, *Politics* and *UK news*, or *Sport* and *Football* categories might be used for the same article in *The Guardian*. Similarly, *Obituary* and *Paid Death Notice*, or *Op-Ed* and *Editorial* in *New York Times* might be used interchangeably as well. Any analysis which focuses specifically on category or genre differences would need to take this into account. For our analysis an approximation of categories is sufficient, since we first need to establish whether there are any differences between the social media popularity models for various categories.

3.2 Social Media Popularity Metrics

Popularity on social media can be measured in a number of ways. We define a news article's popularity as the number of times it is cited on social media, specifically Twitter and Facebook.

Table 3.3: Prevalence of the twenty most frequent categories. Count refers to the absolute count of articles in that category. Percentage refers to the percentage of articles in that category for the dataset.

(a) *The Guardian* corpus

Category	Count	Percentage
World news	4087	15%
Sport	2148	8%
Football	2095	8%
Comment is free	1574	6%
Music	1182	4%
Life and style	1078	4%
Business	1057	4%
UK news	874	3%
Film	772	2%
Media	762	2%
Technology	759	2%
Politics	731	2%
Books	706	2%
Environment	571	2%
Society	554	2%
Stage	493	1%
Television and radio	431	1%
Discover Culture	420	1%
Art and design	405	1%
Money	401	1%

(b) *New York Times* corpus

Genre	Count	Percentage
News	6373	63%
Review	846	8%
Brief	801	7%
Op-Ed	300	2%
Letter	289	2%
Interactive Feature	225	2%
Editorial	198	1%
Schedule	183	1%
Slideshow	177	1%
Obituary	176	1%
List	100	0%
Recipe	88	0%
Paid Death Notice	85	0%
Question	83	0%
Quote	48	0%
Special Report	29	0%
News Analysis	27	0%
Interview	18	0%
Correction	11	0%
Biography	9	0%

3.2.1 Collection and Validation

In order to obtain the number of tweets and retweets, the article's URL was used as the search query for the Twitter Search API[6]. For each article we queried the Twitter API one, three and seven days after the article's publication. An analysis of the resulting popularity measures showed that the measure did not differ significantly after three and seven days, meaning that the propagation through the Twitter network was reached within three days. This follows earlier research by Arapakis et al. (2014) who found that the number of tweets about a news article reaches a peak after approximately two days. This led us to choose **popularity after three days** as the target popularity measures. The collection process was repeated for Facebook likes and shares using the Facebook FQL API[7].

Because the APIs return a sample of all results, we went on to validate the sample we

[6] https://developer.twitter.com/en/docs/tweets/search/overview [Accessed 13th April 2018]

[7] https://developers.facebook.com/docs/graph-api/ [Accessed 13th April 2018]

collected. For a random sample of 100 articles in our dataset we checked the correlations between the citations we collected via the API and the number of citations that appear on *The Guardian* article website. We were unable to repeat this for *New York Times*, as their website does not provide this data for their articles, however if we are able to establish the validity for Twitter sampling for one global news outlet, we expect the validity to carry for another. We report *The Guardian* results in Table 3.4. The correlations are over 0.95 for all measures excluding Twitter after seven days, meaning that our sample is representative.

Table 3.4: Spearman's ρ correlation between Twitter and Facebook popularity using APIs and numbers reported on article website.

	Tweets	Facebook shares
After one day	0.98	0.96
After three days	0.98	0.95
After seven days	0.79	0.96

3.2.2 Popularity Measures

We consider all popularity measures on each social media site jointly (tweets and retweets on Twitter, likes and shares on Facebook), yielding two social media popularity measures:

T = news article's Twitter popularity after three days

F = news article's Facebook popularity after three days

Considering both direct (tweets, shares) and indirect (retweets, likes) citations allows the investigation of the *overall social media popularity* on a given social media site.

Table 3.5 shows the popularity distribution in our datasets. The popularity measures show a highly skewed distribution – most articles achieve a relatively low popularity and there are very few articles that reach very high popularity. This mirrors the results in Arapakis et al. (2014), who found that only 4% of news articles in their dataset reached 100 or more tweets. Twitter and Facebook measures correlate well with each other (0.74 for *The Guardian* and 0.6 for *New York Times*; Spearman's ρ calculated on the training sets). However, Twitter shows a flatter distribution than Facebook. In both datasets the number of citations is much higher for Facebook rather than Twitter, which might be due to the number of users (at the time of data collection Facebook had 1.35 billion active users to Twitter's 0.28 billion according to Statista; a pattern which holds today with 2.07 billion Facebook users to Twitter's 0.33[8]).

[8] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ and

Table 3.5: Headlines corpora summary: number of articles, popularity measure quartiles, and maximum number of citations for any article.

	The Guardian		*New York Times*	
	T	F	T	F
	training: 11,980		training: 5,074	
Q_1	16	11	50	35
Q_2	41	42	102	153
Q_3	88	175	173	739
Max	1,337	122,157	12,738	474,270
	test: 13,806		test: 5,011	
Q_1	19	10	67	24
Q_2	46	47	114	127
Q_3	91	209	188	721
Max	1,287	186,852	6,213	145,118

News source plays an important role, as even two major newspapers like *The Guardian* and *New York Times* show considerable differences in the number of social media citations, where in general *New York Times* articles are more often shared on social media. This follows the earlier findings (Bandari et al., 2012; Arapakis et al., 2014) that news source is the strongest predictor of social media popularity of news articles. However, for a journalist working for a given news outlet, this is a variable they cannot control. In Chapter 8 we build and evaluate prediction models for *The Guardian* and *New York Times* separately, which enables us to draw firmer conclusions about the impact of features on prediction performance without the confounding factor of news source.

Popularity by category. In addition to differences in social media popularity between the two news sources, we also observed differences in social media popularity among the various subsets of the corpora. Tables 3.6 and 3.7 present the most popular categories in the two news corpora. We only considered categories which have at least 10 articles.

The most striking observation for both news sources is that *News* is not the most shared category on social media. Indeed, in *The Guardian* corpus none of the news categories (e.g. *World news*, *UK news*, *Politics*) appear in the top ten most popular categories. Instead, the editorial section (*Comment is free*) and various professional networks are featured.

https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ [Accessed 13th April 2018]

Table 3.6: Top ten most popular categories in *The Guardian* (article count in category must be 10 or above). Count refers to the number of articles in that category. Popularity refers to the median social media popularity for that category. The categories in bold appear in the top ten for both social media popularity measures.

(a) Sorted by median popularity on Twitter.

Category	Count	Popularity
Environment	571	97
Technology	759	97
Teacher Network	74	94
Social Enterprise Network	18	93
Education	279	90
Healthcare Professionals Network	60	85
Science	297	82
Society	554	80
Comment is free	1574	78
Higher Education Network	52	72

(b) Sorted by median popularity on Facebook.

Category	Count	Popularity
Comment is free	1574	232
Environment	571	196
Cities	123	185
Science	297	161
Art and design	405	161
Society	554	144
Global development	192	133
Travel	207	120
Higher Education Network	52	115
Education	279	114

Table 3.7: Top ten most popular genre categories in *New York Times* (article count in category must be 10 or above). Count refers to the number of articles in that category. Popularity refers to the median social media popularity for that category. The categories in bold appear in the top ten for both social media popularity measures.

(a) Sorted by median popularity on Twitter.

Category	Count	Popularity
News Analysis	27	214
Op-Ed	300	156
Question	83	144
Obituary	176	139
News	6373	129
Editorial	198	119
Interview	18	113
Review	846	86
Special Report	29	83
Schedule	183	71

(b) Sorted by median popularity on Facebook.

Category	Count	Popularity
Op-Ed	300	1492
News Analysis	27	803
Obituary	176	569
Editorial	198	501
News	6373	222
Question	83	188
Review	846	178
Special Report	29	139
Interview	18	114
Recipe	88	34

Similarly, in *New York Times* corpus editorials (*News Analysis* and *Op-Ed*) are the top of the list, and *News* only appears in fifth place. In terms of differences between social media popularity measures, there is not much variety in *New York Times* results. This is however due to a relatively small number of genres (27 unique genres). In the case of *The Guardian*, six out of ten most popular categories overlap. Interestingly, four professional network categories are in the top ten for Twitter, suggesting high level of engagement on social media among these communities. These differences in the popularity of categories can influence the impact of individual features (e.g. certain features having a much greater/lesser influence on social media popularity in particular categories) and the resulting prediction model for a particular category. We investigate such a scenario in Section 8.2.3.

3.3 Summary

In our experimental approach to modelling social media popularity of news articles using headlines we want to extract news values and style features from headline text and investigate their impact on social media popularity. In this chapter we described the data collection process for headlines corpora from two global broadsheet news sources – *The Guardian* and *New York Times*. We also described the process of collecting and validating the associated social media popularity of news articles on Twitter and Facebook. The datasets we created are made publicly available (cf. Appendix A).

Through an overview of the datasets we created we identified issues arising from the data, most notably the skewed distribution of popularity and differences in the prevalence and popularity of news article categories. We need to account for these when evaluating features and building prediction models. In the next chapter we turn to the implementation and evaluation for the first feature group we use – news values.

Chapter 4

Implementing News Values

Our goal is to model social media popularity of news articles using headlines. In Chapter 2 we identified several research gaps relating to this goal: (i) only a small number of text characteristics have been used in news article popularity prediction models (Section 2.2), (ii) headlines play a crucial function in digital spaces, however these prediction models have only used full article text (Section 2.3), and (iii) news values offer a valuable journalistic perspective, but no computational extraction methods are available (Section 2.5).

In this chapter we propose a new perspective for analysing the content of headlines – news values. We present the first fully *automatic and topic-independent extraction methods of news values from headline text*. Our operationsalisation of news values relies on explicit and topic-independent linguistic indicators of news values which are captured using state-of-the-art NLP methods. This approach has the benefit of not having to develop or reimplement methods to suit new domains or topics (cf. this type of adjustment would need to be made for a semi-automatic approach to news values extraction like in Potts et al. (2015)). We are the first to propose using Magnitude, Proximity, Surprise, and Uniqueness for the task of modelling social media popularity of news articles. We also add to the features that have been used previously (we add wikification and burstiness to Prominence, and connotations and bias to Sentiment).

4.1 Implementation

The following sections provide the details of an automatic extraction of six news values from headlines: Prominence, Sentiment, Magnitude, Proximity, Surprise, and Uniqueness. Justification of our choice and an overview of these news values was presented in Section 2.5. In Table 4.1 we present a summary of the feature implementations, as well as general statistics about the occurrence of news values in two headlines corpora: *The Guardian* and *New York Times*. In the following sections for each news value we give a justification and implementation of our operationalisation and provide some examples of their application on headlines corpora. Examples of *The Guardian* and *New York Times* headlines annotated with news values are included in Appendix C.

4.1.1 Prominence

Prominence can be interpreted as eliteness, or recognisability (cf. Section 2.5.2). We approximate prominence as the amount of online attention an entity gets. More online attention indicates popularity and/or recognisability (e.g. average number of daily Wikipedia pageviews distinguishes between three bands of varying popularity: 8248 pageviews for One Direction, 1054 for X Ambassadors, and only 10 for Warsaw Village Band[1]). Our implementation of Prominence is the first to use two state-of-the-art techniques for the task of news article popularity prediction: wikification and burstiness. Firstly, because of the social media aspect of the prediction model, we adopt a broad definition of entity to identify entities in headline text. Wikification (e.g. Mihalcea and Csomai (2007)) is a method of linking keywords in text to a relevant Wikipedia page. Using wikification means considering not only what has traditionally been seen as entities (people, organisations, locations), but also concepts, titles of books, TV shows, films, etc. Moreover, when we compared TagMe wikification against a traditional entity recognition tool (Stanford Named Entity Recognizer (Finkel et al., 2005)), using wikified entities yielded more highly correlated results with social media popularity measures (statistically significant at p<0.05, calculated on the training sets). Secondly, as online prominence varies with time, we consider several temporal aspects: long-term vs. recent prominence and burstiness. We are the first to consider the burstiness of an entity's prominence in news article popularity prediction.

We implement six Prominence features, following three steps.

Firstly, we use the **number of entities** in the headline as identified by TagMe. For example, the TagMe output for *The Guardian* Headline 4.1 yields four entities: WAR

[1] http://bit.ly/2AUrQoA [Accessed 13th April 2018]

Table 4.1: Summary of news values feature implementations. Notation is explained in Section 3.1.2. Examples of annotated headlines are given in Appendix C.

Feature name	**Implementation**
Prominence	
N1: number of entities	$\lvert E \rvert$
N2: Wikipedia long-term prominence	$\sum_{e \in E} pageviews_{e,d_{-365},d_{-1}}$
N3: Wikipedia day-before prominence	$\sum_{e \in E} pageviews_{e,d_{-1},d_{-1}}$
N4: news source recent prominence	$\sum_{e \in E} newsmentions_{e,d_{-7},d_{-1}}$
N5: Wikipedia current burst size	$\sum_{e \in E} daysburst_{e,d_{-1},d_{-1}} \times \frac{pageviews(e,d_{-1},d_{-1}) - \text{mean}(MA_e)}{\text{SD}(MA_e)}$
N6: Wikipedia burstiness	$\sum_{e \in E} daysburst_{e,d_{-365},d_{-1}}$
Sentiment	
N7: sentiment	$max\_positivity - max\_negativity - 2$
N8: polarity	$max\_positivity + max\_negativity$
N9: connotations	$\frac{\text{\# content words with positive or negative connotations}}{\lvert C \rvert}$
N10: bias	$\frac{\text{\# biased content words}}{\lvert C \rvert}$
Magnitude	
N11: comparative/superlative	$\frac{\text{\# words with JJR\|JJS\|RBR\|RBS POS tag}}{\lvert C \rvert}$
N12: intensifiers	$\frac{\text{\# intensifiers}}{\lvert H \rvert}$
N13: downtoners	$\frac{\text{\# downtoners}}{\lvert H \rvert}$
Proximity	
N14: proximity	1 if explicit reference to UK/US in H or in Wikipedia category tags, else 0
Surprise	
N15: surprise	$min\text{LL}_p$ where LL_p is the log-likelihood for a phrase in H
Uniqueness	
N16: uniqueness	$max_{t \in d-72hr}$ cosine similarity$(H, pastH_t)$

HORSE (FILM), JEREMY IRVINE, GAY, and STONEWALL RIOTS. By using wikification we obtain a wider variety of entities – not just the person (JEREMY IRVINE), but also a film title (WAR HORSE (FILM); in this context actually a play), a concept (GAY), and an event (STONEWALL RIOTS). Headline 4.2 provides an interesting example of TagMe output. Even though the output UNIVERSITY OF MONTANA is a mistake in entity linking ('Montana' should be linked to MONTANA as in state), it is notable that TagMe chose this entity, because of the token *Professors*. While in this case TagMe made a mistake, generally the contextualisation of entity links is an advantage.

Headline 4.1. "War Horse's Jeremy Irvine to star in gay rights film Stonewall"

Headline 4.2. "Professors' Research Project Stirs Political Outrage in Montana"

Secondly, we implement three features which aggregate prominence over certain time periods. For an entity e, we denote as $pageviews_{e,d_{-m},d_{-n}}$ the median number of Wikipedia daily page views[2] for that entity between days d_{-m} and d_{-n}. Day numbering is determined in reference to the article publication day d. **Wikipedia long-term prominence** is calculated over one year ($pageviews_{e,d_{-365},d_{-1}}$), and **Wikipedia day-before prominence** on the day before publication ($pageviews_{e,d_{-1},d_{-1}}$). Following experimentation, we found the previous day's prominence to be closest to the actual on-the-day prominence[3]. For a news-centric perspective of prominence, we also calculate **news source recent prominence**: the sum of e's mentions in the news source headlines in the week before publication day, denoted as $newsmentions_{e,d_{-7},d_{-1}}$.. Using Wikipedia as a prominence source allows us to differentiate between entities. For example, using examples from Appendix C compare the low values for long-term and day-before prominence for Headline 4.3 to the high values in Headline 4.4. Considering both long-term and day-before prominence also allows to differentiate between entities and identify headlines which mention important entities at the time. For example, Headline 4.5 has three quite prominent entities (FC BAYERN MUNICH, A.S. ROMA, UEFA CHAMPIONS LEAGUE; long-term prominence = 13,793), but at that particular time these entities attracted even more attention on Wikipedia (day-before prominence = 26,993). Another example of a similar phenomenon which also exemplifies the use of recent news source prominence is Headline 4.6, where prominence goes from long-term = 7284 to day-before = 12,108 and news source prominence = 7 (when the median for that feature is 0, cf. Table 4.2). This highlights the need to implement a variety

[2] http://dumps.wikimedia.org/other/pagecounts-ez/ [Accessed 13th April 2018]

[3] We compared the on-the-day prominence with: (1) Holt's linear model, (2) Holt-Winters seasonal additive model, (3) median prominence over one year, (4) median prominence over seven days, and (5) prominence on the day before publication using Spearman's ρ and mean absolute scaled error. For both measures the day-before publication method had best results (ρ=0.96 and MASE=0.41).

of features, so that nuances like the aforementioned examples can be modelled.

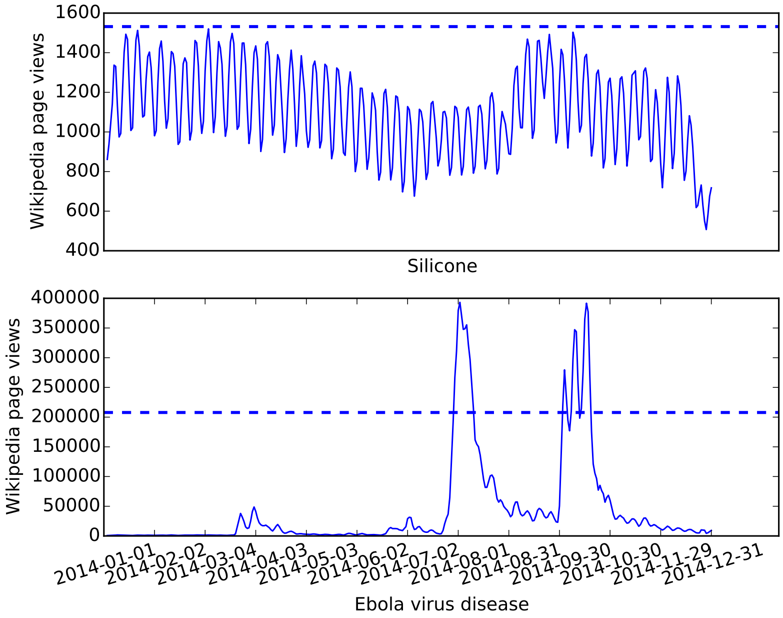
Headline 4.3. "Getting creative work done with OLIVER BURKEMAN" (long-term prominence = 14, day-before prominence = 30)

Headline 4.4. "MANCHESTER CITY recover poise before final straight in title chase" (long-term prominence = 5331, day-before prominence = 4487)

Headline 4.5. "BAYERN MUNICH Cruises Past ROMA in CHAMPIONS LEAGUE" (long-term prominence = 13793, day-before prominence = 26993)

Headline 4.6. "SOUTH KOREA ferry disaster: footage shows crew being rescued" (long-term prominence = 7284, day-before prominence = 12108, news source prominence = 7)

Figure 4.1: Time series plots of Wikipedia page views moving averages (MA) for two entities: non-bursty SILICONE (top) and bursty EBOLA VIRUS DISEASE (bottom). The dashed line shows a global burst cut-off line.



Thirdly, as entities exhibit different temporal patterns of prominence, we differentiate between entities which have a *steady* prominence (e.g. SILICONE) and entities which become *bursty*, i.e. suddenly prominent for a short period of time (e.g. EBOLA VIRUS). We are the first to consider burstiness for popularity prediction of news articles. To identify bursty entities, we implement the burst detection algorithm by Vlachos et al. (2004) (cf. Algorithm 1). An entity is defined as *being in a burst* if its moving average in a given time frame is above the cut-off point (cf. Figure 4.1). We use entity bursts in two ways. Firstly, **current burst size** indicates how many standard deviations above MA_e is any e which is in a burst day before publication ($daysburst_{e,d-1,d-1}$ returns 1 if e is in a burst, 0 if not). An example of an entity in a burst is provided in Headline 4.7. Secondly, **burstiness** indicates the number of days that e was in a burst over a year ($daysburst_{e,d-365,d_{-1}}$). This feature points to entities that regularly become bursty, but might not be in a burst when the headline is published. For example, Headline 4.8 has three entities (BROOKLYN, DISTRICT ATTORNEY, MURDER), and while none of them are in a burst, their total burstiness is 56 (compared to *New York Times* median for that feature which is 15; cf. Table 4.2). These features provide another dimension to exploring the effect of Prominence which has not been considered before in research on news article popularity prediction.

Algorithm 1 Burst detection algorithm adapted from Vlachos et al. (2004). Following experimentation, we set the moving average size to three days and the cut-off point to two times standard deviation.

```
1: Calculate moving average of length 3 for entity $e$ ($MA_e$) for sequence $d_{-365}, ...d_{-1}$.
2: Set cutoff = $mean(MA_e) + 2 \times SD(MA_e)$
3: Bursts = $d_i | MA_e(i) > cut\text{-}off$
```

Headline 4.7. "Teenage plane STOWAWAY snuck aboard despite being caught on camera" (burst size = 41.7; long-term prominence = 182, day-before prominence = 14259)

Headline 4.8. "BROOKLYN DISTRICT ATTORNEY Will Ask Judge to Throw Out MURDER Convictions" (burstiness = 56, current burst size = 0)

Finally, we note that as one headline can refer to more than one entity, all prominence measures are aggregated into headline features via summation over all entities in the headline (see Table 4.1).

4.1.2 Sentiment

We look at both direct and indirect expressions of Sentiment in line with journalism literature (cf. Section 2.5.2). As direct measures, we combine SentiWordNet (Baccianella et al., 2010) positivity and negativity scores of a headline's content words, and calculate both **sentiment** and **polarity** scores following Kucuktunc et al. (2012). Sentiment values lower than -2 indicate more negative sentiment, cf. Headline 4.9. Polarity indicates the overall strength of the sentiment in the headline, cf. Headline 4.10. Sentiment can also be indirect. To explore this we first looked at **connotations**, whereby a word may be in itself objective, but carry a negative connotation (e.g. *scream*). We therefore measure the proportion of content words in a headline with a positive or negative connotation, as indicated in a connotations lexicon (Feng et al., 2013). Headline 4.11 is an example of a high connotations score but neutral sentiment according to direct measures of Sentiment. Secondly, we measure the proportion of **biased** content words. For example, the same political organisation can be described as *far-right*, *nationalist*, or *fascist*, each of these words indicating a bias towards a certain – often subjective – reading. We used a bias lexicon by Recasens et al. (2013) to obtain a list of biased words. For example, Headline 4.12 has a high bias score at 0.3. The biased words in this headline are: *international*, *rights*, *abuse*. As with the previous example, the direct measures of Sentiment for this headline indicate a neutral headline, but by adding features which target indirect sentiment we are able to explore another dimension of Sentiment for headlines.

Headline 4.9. "Faces of Breast Cancer" (sentiment=-2.25)

Headline 4.10. "Martin Kaymer the Masters survivor driven by high ambition of Ryder Cup" (polarity=0.75))

Headline 4.11. "As Egyptians Grasp for Stability, Sisi Fortifies His Presidency" (sentiment neutral at -2, polarity = 0, connotations = 0.67)

Headline 4.12. "Coalition Seeks to Send North Korea to International Court Over Rights Abuses" (bias = 0.3)

4.1.3 Magnitude

We focus on explicit linguistic indicators of event size (comparatives/superlatives and amplifiers) which provides us with a topic-independent implementation. For example, the intensifier *many* indicates a high number regardless of the actual event size which is relative (e.g. "many victims" would refer to a higher number for an earthquake than a car accident).

The implementation includes the proportion of **comparative and superlative** adjectives and adverbs (indicated by part-of-speech tags obtained in the preprocessing stage; cf. Headline 4.13). We also use two types of amplifiers: intensifiers and downtoners. **Intensifiers** are used to enhance or heighten the lexical item they describe (e.g. *extremely* dangerous). Conversely, **downtoners** are used to diminish the scope of the lexical items they describes (e.g. *almost* there). To obtain the proportions of amplifiers (intensifiers and downtoners separately) we combined the vocabulary lists in Quirk et al. (1985, p.589, 597) and Biber (1991, p.240), which resulted in wordlists of 248 intensifiers and 39 downtoners. These wordlists have been made publicly available[4].

[4] https://apiotrkowicz.wordpress.com/datasets/

Headline 4.13. "*Latest* Alaska Polls Show Surprising Shift Toward Mark Begich"

4.1.4 Proximity

Our implementation of Proximity indicates whether a headline refers to an entity that is geographically close to the news source. We chose geographic rather than cultural proximity, because our approach uses explicit indications of news values and geographic locations are indicated explicitly in the language, while culture would require an analysis which considers world knowledge. Proximity to the news source is our approximation for proximity to the reader – we assume that most readership of *The Guardian* is British and most readership of *New York Times* is American. To identify explicit indicators of proximity to United Kingdom (for *The Guardian*) and Unites States (for *New York Times*) we manually create a wordlist of country-specific terms, including names and variations for the country, regions, capital city. This resulted in 17 UK-related terms for *The Guardian*, and 61 US-related terms for *New York Times* (which have been made publicly available[5].). We then look for matches both in the headline text (e.g. Headline 4.14), as well as in the names of Wikipedia categories of each entity supplied in the TagMe output (e.g. category POSTAL SYSTEM OF THE UNITED KINGDOM for Headline 4.15). The feature is binary indicating whether or not there is an entity related to UK or US in the headline.

Headline 4.14. "*London* smog warning as Saharan sand sweeps southern *England*"

Headline 4.15. "Undervaluing *Royal Mail* shares cost taxpayers £750m in one day"

4.1.5 Surprise

In our implementation of Surprise we target explicit surprise which arises through surprising language (cf. Section 2.5.2 for distinction between implicit and explicit surprise). We identify surprising language by looking at selectional preferences of lexical items. We do this by calculating the commonness of phrases in headlines with reference to a large corpus. We broadly adapt the method used by Louis and Nenkova (2013) to identify creative word pairs in popular science writing. We first extract phrases of following types: SUBJ-V, V-OBJ, ADV-V, ADJ-N, N-N (where N is noun, V is verb, ADJ is adjective, and ADV is adverb). These cover the most meaningful pairs, i.e. descriptors of nouns and verbs. We then generate a regular expression with their inflected forms (e.g. *man drinks* → *man drinks|drank|drinking*), which provides us with the linguistic variations for a phrase. For each regular expression we obtain its frequency in a large publicly available Wikipedia

[5] https://apiotrkowicz.wordpress.com/datasets/

corpus[6]. We found Wikipedia to yield better results than the Google Ngrams corpus; perhaps because there is less noise which gives a more accurate reflection of surprising phrasing. The frequencies are summed for each phrase and the log-likelihood (LL) for the phrase is calculated. The lower the log-likelihood, the more surprising the phrase. The feature value equals the lowest log-likelihood in the headline, as we are looking for the most surprising phrase.

4.1.6 Uniqueness

An analysis of several storylines in our headlines datasets showed that of two very similar headlines, the latter tends to be less popular. For example, during a developing story on a South Korean ferry sinking in April 2014 in *The Guardian* we noticed pairs of very similar headlines, where the later ones had a much lower popularity on social media (see pairs: Headlines 4.16 and 4.17, and Headlines 4.18 and 4.19), which led us to investigate whether having highly similar headlines published in close temporal proximity had an effect on social media popularity.

Headline 4.16. "South Korea ferry captain arrested" (Twitter popularity T: 216)

Headline 4.17. "South Korea ferry: captain arrested after sinking – video" (T: 18)

Headline 4.18. "Ferry disaster: South Korean prime minister resigns" (T: 99)

Headline 4.19. "South Korean prime minister resigns over ferry sinking" (T: 30)

First, we need to establish the time lapse between two headlines being published which would trigger this decrease in popularity due to similarity. We experimented with a number of cut-off points and found that 72 hours yielded best results. We also noted that storylines often share entities (e.g. SOUTH KOREA in the examples above). We found that having entity overlap helps with ensuring that the headlines are part of the same storyline, while including headlines with no entities ensures more coverage of the dataset. To calculate Uniqueness for a given headline H we select past headlines from 72 hours before its publication and which have at least one TagMe entity overlapping or neither has any entities. For a pair of H and $pastH$ vectors (created using a *tf-idf* weighted Gigaword corpus) we calculate their cosine similarity. The highest cosine similarity is assigned as the feature value, meaning that the higher the value, the more similar are the headlines.

[6] http://nlp.cs.nyu.edu/wikipedia-data/ [Accessed 13th April 2018]

Table 4.2: News values feature statistics on *The Guardian* and *New York Times* corpora. Reported measures: median and maximum values, prevalence (percentage of non-zero scores).

Feature name	*The Guardian*			*New York Times*		
	Med.	Max.	Prev.	Med.	Max.	Prev.
number of entities	1	8	79%	1	4	100%
Wikipedia long-term prominence	1342	125757	79%	626	65,327	66%
Wikipedia day-before prominence	1642	1031722	78%	773	467,458	66%
news source recent prominence	0	122	50%	0	70	32%
Wikipedia current burst size	0	57.16	12%	0	57.18	10%
Wikipedia burstiness	21	156	78%	15	166	66%
sentiment	-2	-1	100%	−2	−1	100%
polarity	0.5	1.88	79%	0	1.88	43%
connotations	0.34	1	92%	0.25	1	78%
bias	0.13	1	61%	0.11	1	51%
comparative/superlative	0	1	7%	0	1	3%
intensifiers	0	0.34	10%	0	0.33	6%
downtoners	0	0.29	4%	0	0.33	3%
proximity	0	1	35%	0	1	32%
surprise	4.15	2726186	100%	4.04	2724886	100%
uniqueness	0	0.83	13%	0	1	34%

4.2 Evaluation

To evaluate the news values feature extraction methods presented in the preceding sections, we applied them on two corpora consisting of headlines from *The Guardian* and *New York Times* (cf. Chapter 3 for details of data collection and preprocessing). These sources provide a wide coverage of various topics and genres, allowing a good exploration of news values, as well as a comparison between British and American English news sources. We also calculate some general occurrence statistic of the extracted features for each corpus, including median, maximum and prevalence (ratio of non-zero scores) which are reported in Table 4.2. It shows that there are differences in how often the news values features occur in the headlines corpora (prevalence) and what are the average values for the features (median). Prominence, Sentiment, and Surprise occur commonly in the corpora and have non-zero median values, indicating a wide variety in feature values which can be helpful for the prediction models we present in Chapter 8. Some of the news values we propose (Magnitude, Proximity, Uniqueness) are less common and their medians are zero. Although they are not very common in our corpora, this might not be the case for other news sources. In Chapters 6 and 7 we investigate the impact of these features on social media popularity and perceived popularity, which provides us insights into how a headline can be reformulated to reach higher popularity on social media. The corpus statistics for each feature are discussed in depth in Section 4.2.4. We now turn to the evaluation of news values extraction methods which we conduct by comparing the automatic extraction of news values to a manually annotated gold standard.

4.2.1 Manually Annotated Gold Standard

For each news value we select 20 headlines from *The Guardian* headlines corpus. In order to obtain annotations which reflect the variety of possible feature values in the corpora we need to select representative examples of headlines. To do that we randomly select 10 headlines from the top quartile feature values and 10 from the bottom quartile. Quartiles are computed using the feature values for a given news value (cf. Table 4.2 for a summary). For news values that are split into multiple features (Prominence, Sentiment, Magnitude), the feature group vectors are ordered to obtain quartiles. Overall, a total of 120 headlines were selected for manual annotation. We obtain the manual gold standard separately for the first five news values (Prominence, Sentiment, Magnitude, Proximity, Surprise) and for the news value of Uniqueness. That is because comparing headlines is required to annotate the news value of Uniqueness, while for the first five news values only the given headline is needed. For the first five news values we had three expert annotators, PhD students in linguistics, annotate each headline as positive or negative (Y/N). For the news value of Uniqueness, the annotators were presented with 20 headlines from the corpus and for each headline another 20 past headlines for comparison with highest and lowest headline uniqueness scores (which had been randomly sampled). The annotators indicated whether any of the past headlines were very similar (i.e. highly related) to a given headline. Table 4.3 gives examples of headlines with the manual annotations labels and the automatically extracted feature values.

4.2.2 Inter-Annotator Agreement

The inter-annotator agreement (IAA) was calculated using Fleiss' kappa (Fleiss, 1971). Fleiss' kappa (κ) allows to calculate inter-annotator agreement for cases where the data is categorical and there are more than two annotators. The IAA results are reported in

Table 4.3: Examples of annotated headlines. Y/N indicate a yes/no majority vote manual annotation. Below are the automatically extracted values aggregated by feature group (cf. Table 4.2 for feature value ranges).

#	Headline	Prom.	Sent.	Magn.	Prox.	Surp.
E1	"Getting really hung up on EE/Orange customer service"	Y 0	Y 3	Y 0.125	Y 0	Y 3.23
E2	"Mount Everest avalanche leaves at least 12 Nepalese climbers dead"	Y 13272	Y 4.25	Y 0.17	N 0	N 4.15
E3	"Huzzah for foreign experts. After all, they're better than our own"	N 672	Y 2.75	Y 0.2	N 0	Y 398
E4	"Rev; Martin Amis's England; and A Very British Renaissance: TV review – video"	Y 36236	N 2.45	N 0.08	Y 1	N 4.15
E5	"This week's new live comedy'	N 0	N 3.25	N 0	N 0	N 102

Table 4.4. The inter-annotator agreement ranges from substantial for Prominence (.76) and Uniqueness (0.73), through moderate for Magnitude (0.43), Surprise (0.48), and Proximity (0.55), to fair for Sentiment (0.22). The inter-annotator agreement for each news value is discussed in depth in Section 4.2.4. Some news values (Sentiment, Magnitude, Surprise) were found to be particularly challenging for humans to agree on, which suggests that the perception of these news values is very subjective. For example, annotators disagreed about sentiment in Headline 4.20. Similarly, annotators disagreed about Magnitude in Headline 4.21, and about Surprise in Headline 4.22. The annotators themselves remarked that sometimes they chose 'on instinct' and that their responses might vary from day to day. This highlights the challenging nature of the task of automatic detection of news values, as news values are somehow tacitly understood. The annotators' judgments were aggregated using a majority vote, resulting in the gold standard for evaluation.

Headline 4.20. "Unthinkable? A rethink of classic remakes"

Headline 4.21. "Spain's wetlands wonder is under threat for a second time in 16 years"

Headline 4.22. "The Amazing Spider-Man 2 review: 'so savvy, punchy and dashing that it won't be denied"'

Table 4.4: Inter-annotator agreement scores for manual annotation of news values for *The Guardian* corpus (N=3).

News value	**Fleiss' κ**	**Agreement level**
Prominence	0.76	Substantial
Sentiment	0.22	Fair
Magnitude	0.43	Moderate
Proximity	0.55	Moderate
Surprise	0.48	Moderate
Uniqueness	0.73	Substantial

4.2.3 Comparison with Gold Standard

We calculate pairwise comparisons between each feature value and the relevant manual label (e.g. number of entities and Prominence, bias and Sentiment; cf. Table 4.5). The Kruskal-Wallis test (Kruskal and Wallis, 1952) is used to determine whether the differences in feature values for the two manual annotation labels (Y/N) are significant (cf. Table 4.5). These results indicate whether the value calculated for a given feature correctly reflects the presence of a news value in the gold standard produced by the human annotators (cf. examples of headlines with both the manual label and the extracted feature values in Table 4.3). The results of the evaluation for each news value are discussed in the next section.

4.2.4 Discussion

The evaluation of the feature extraction methods for each news value covers three aspects: (i) application on headlines corpora (refers to the feature implementation statistics in Table 4.2), (ii) results of the human annotation task (refers to the inter-annotator agreement scores in Table 4.4), and (iii) the appropriateness of the feature extraction methods (refers to the Kruskall-Wallis test results of comparing the automatic extraction scores to the manual annotation labels in Table 4.5).

Prominence

Application on headlines corpora. It occurs quite frequently – most headlines in *The Guardian* corpus have at least one entity (median number of entities = 1), which attracts a fair amount of online attention (median Wikipedia long-term prominence = 1,342 pageviews). Some headlines include very prominent entities (maximum Wikipedia day-before prominence = 1,031,722). The outputs from *New York Times* are similar – every headline is associated with at least one Wikipedia entity (100% prevalence for number of entities); and Wikipedia burstiness, long-term, and day-before prominence have non-zero scores in 66% of headlines. This shows that Wikipedia provides a wide coverage for

Table 4.5: Comparison of automatic news values extraction against manual gold standard. Reported are: the results of the Kruskall-Wallis test comparing the manual gold standard label to computationally extracted feature value for *The Guardian* corpus (* *p*<0.05, ** *p*<0.01, *** *p*<0.001), and the gold standard category that was used for the comparison.

Feature name	**Kruskall-Wallis test**	**Gold standard category**
number of features	***	Prominence
Wikipedia long-term prominence	***	
Wikipedia day-before prominence	***	
news source recent prominence	**	
Wikipedia current burst size	0.2	
Wikipedia burstiness	***	
sentiment	0.1	Sentiment
polarity	**	
connotations	0.2	
bias	*	
comparative/superlative	***	Magnitude
intensifiers	***	
downtoners	0.2	
proximity	***	Proximity
surprise	*	Surprise
uniqueness	*	Uniqueness

the computation of Prominence. Wikipedia current burst is a rare feature (12% in *The Guardian* and 10% in *New York Times*), as capturing an entity in a burst is uncommon, since bursts do not apply to all entities and do not happen frequently. A possible explanation is that news readers do not immediately turn to Wikipedia to look up entities currently in the news (which would decrease the burst size), or bursts might develop for longer than a day.

Human annotation. The inter-annotator agreement is the highest for this news value (κ=.76) and indicates substantial agreement. This suggests that Prominence is perceived similarly by human annotators.

Appropriateness of feature extraction. Nearly all Prominence features reach significance level of p<0.001 when compared to the manual annotations. This strongly supports our implementation of Prominence, in particular the use of wikification and Wikipedia as a prominence source. Burstiness, which uses a burst detection algorithm tailored for our specific task, presents a new way of looking at Prominence. While burstiness (i.e. how many times in a year an entity had page views significantly higher than its average) is a reliable feature, current burst size (i.e. size of the burst on the day before article publication) is not significantly associated with the gold standard.

Overall, Prominence is one of the most prevalent news values and our approach using TagMe wikification proves very reliable.

Sentiment

Application on headlines corpora. Headlines in the broadsheet newspapers we consider tend to be quite neutral (*The Guardian*: median sentiment = -2 and median polarity = 0.5; *New York Times*: sentiment = -2 and polarity = 0). However, most headlines in *The Guardian* contain at least some connotations or bias (connotations prevalence = 92%, bias prevalence = 61%; slightly lower in *New York Times*: 78% and 51%). This supports our decision to look at indirect Sentiment (connotations and biased language), since they appear frequently in the headlines corpora.

Human annotation. The inter-annotator agreement was fair, at κ=.22. The fact that many headlines are neutral can explain the low agreement, since the neutral cases are where experts are more likely to disagree (e.g. Headline 4.23 which reports a business event, but it can interpreted as using sentiment-charged language through phrases *bring in new* and *profit slide*). Sentiment was also noted to be quite subtle, e.g. Headline 4.24, where depending on the point of view the headline can be interpreted as neutral reporting, or a subtle jibe. Furthermore, reported IAA scores for Sentiment vary in literature. For example, manual annotation for one aspect of Sentiment like positivity/negativity can achieve substantial agreement (e.g. 0.76 agreement between experts in Snow et al. (2008)), but when looking specifically at headlines di Buono et al. (2017) report κ=0.47 for *Bad news* and κ=0.23 for *Good news*, which indicates that human annotators find Sentiment annotation in headlines challenging.

Headline 4.23. "Debenhams to bring in new names after profit slide"

Headline 4.24. "Paul Sykes: the man spending millions to make Ukip's posters visible from space"

Appropriateness of feature extraction. When compared to the manual annotations, two (polarity and bias) out of four Sentiment features reach significance levels, so our implementation does capture some aspects of Sentiment. Extracting Sentiment from headlines proves a challenge, since they are short texts with limited context and often the sentiment is implied or requires world knowledge to identify (e.g. “Guinea’s Ebola outbreak: what is the virus and what’s being done?”).

Overall, operationalising Sentiment for headline text highlighted a number of issues. Firstly, it is not typical for the broadsheet headlines corpora we use. For example, Reis et al. (2015) found that over 40% of headlines in two ‘broadsheet’ news sources were neutral. Secondly, the interpretation of Sentiment in headlines can be very subjective. We noted that sentiment-charged language in headlines does not always accurately reflect the true sentiment or emotion of the author and/or reader. On one hand, there are highly evocative headlines that describe some tragic news events (+sentiment, +emotion). On the other hand, there are headlines which use sentiment-charged language, but are not evocative to the same extent (+sentiment, -emotion). For example, *comedy* in example E5 in Table 4.3 has positive sentiment, but does not evoke positive emotion. Provided it can be done reliably, disentangling sentiment and emotion might paint a clearer picture. However, this would require classification along two different axes: firstly, sentiment (positive, neutral, negative), and secondly, emotion (anger, disgust, fear, happiness, sadness, and surprise (Ekman, 1992)); and headlines might not provide enough context for a reliable emotion classification.

Magnitude

Application on headlines corpora. It is the least prevalent news value (between 4-10% for all three features in *The Guardian*; between 3-6% in *New York Times*). The median values are also all zero. Our implementation of this news value could be the reason for this, since we utilise only explicit and topic-independent linguistic indicators of Magnitude (e.g. *hardly*, *highest*) in order to facilitate extending the approach to other domains. In actual news headlines, there are often implicit indicators (e.g. “20 people killed” implying that this is a significant number), which are harder to capture reliably using computational methods, as they require the coding of world knowledge.

Human annotation. The IAA was moderate (κ=.43), which suggests that the annotators found it difficult to agree whether the scope of a headline is considerably greater/smaller. This indicates that the perception of Magnitude in headlines is subjective.

Appropriateness of feature extraction. Two out of three features were significant at p<0.001. This confirms that our approach that relies on part-of-speech tags and wordlists captures this news value. The only feature not to reach a significance level was downtoners. Downtoners are a class of words which aim to diminish the word they describe (e.g. *nearly*, *barely*, *just*). They are not only rare (prevalence is 4% in *The Guardian*, 3% in *New York Times*), but also require specific knowledge to identify them (we identified 39 downtoners, compared to 248 intensifiers). Bearing in mind that downtoners might have more impact if their coverage increases with a more comprehensive wordlist, the other Magnitude features (comparative/superlative and intensifiers) can be reliably used for headlines.

Overall, Magnitude is rare and seems to be judged subjectively by annotators, but we reliably extracted it using comparatives/superlatives and intensifiers.

Proximity

Application on headlines corpora. This news value occurs in only a third of headlines in our corpora (35% of *The Guardian* headlines, and 32% of headlines in *New York Times*). This is not surprising, considering that both *The Guardian* and *New York Times* have a global readership, so the majority of headlines does not relate to UK/US. We would expect similar prevalence results for other global news outlets, while for regional news outlets (e.g. *Yorkshire Evening Post*) Proximity might be more common.

Human annotation. The IAA is moderate (κ=.55). The annotation of Proximity is dependent on entities which indicate relation to a certain country to be recognised by annotators, however we find that familiarity with entities can vary considerably (cf. Section 9.2.3).

Appropriateness of feature extraction. The feature reaches significance at p<0.001, so our method of capturing Proximity with a wordlist and entity categories is well-supported. Using entity categories ensures wider coverage and less manual effort than just using a wordlist. This is turn depends on the reliability of the named-entity recogniser or wikification tool. In some cases an entity might be missed (cf. example E1 in Table 4.3, where *EE/Orange* has been missed and consequently both Prominence and Proximity scores are zero). It is important to note that the news value of Proximity covers both geographic and cultural proximity. In our evaluation the annotators were UK residents, familiar with *The Guardian*, but demographics of the reader will probably influence their familiarity with some entities (this is addressed in Chapter 9).

Overall, Proximity is not frequent, but our approach using a wordlist and Wikipedia categories proves reliable.

Surprise

Application on headlines corpora. The median log-likelihood for this features is relatively low (4.15 for *The Guardian* and 4.04 for *New York Times*), which means that most headlines have fairly surprising phrasing. This might be because headlines do not tend to strictly follow the conventions of everyday language (e.g. frequent use of untensed verbs and noun clusters). Using a corpus which has not been tailored to a domain (we used

Wikipedia) might result in lower log-likelihood. However, because of rate limiting of news APIs, creating a large enough corpus of current headlines would take a prohibitively large amount of time.

Human annotation. The IAA was moderate at κ=.48. As is the case with Sentiment and Magnitude, the moderate agreement for Surprise can be due the subjectivity in how this news value is perceived.

Appropriateness of feature extraction. The feature is significant (p<0.05). This shows that using shallow count-based methods can reliably capture this news value. In other genres where Surprise might play a bigger role, this method can be extended by using a headline-specific corpus or building language model that takes into account syntactic structure.

Overall, our approach which targets surprising phrasing using a Wikipedia-based language model captures the news value of Surprise.

Uniqueness

Application on headlines corpora. The prevalence is quite low (15% for *The Guardian*, but slightly higher at 34% in *New York Times*), which follows the basic journalistic principle that news has to add something new, i.e. headlines at a given time should be dissimilar to the ones that were published recently.

Human annotation. IAA was substantial with κ=.73, which indicates that annotators generally agree about headline similarity.

Appropriateness of feature extraction. The feature was significant (p<0.05), so we can be sure that any similar headlines are identified. An analysis of headlines with non-zero Uniqueness values reveals that most of them are either part of a regular feature (e.g. "Reviews roundup"), or part of continuing storylines about the same event (often featuring some additional reporting which uses diffetent media like video or picture gallery).

Overall, headlines which do not include the news value of Uniqueness are fairly rare, but our implementation reliably identifies such instances.

4.3 Summary

In this thesis we model the social media popularity of news articles using news values and style features of headlines. In this chapter we presented the operationalisation of six news values features which we extract computationally from headline text: Prominence, Sentiment, Magnitude, Proximity, Surprise, and Uniqueness. The implementations used explicit and topic-independent linguistic indicators in order to make our methods more generalisable. We evaluated the news values operationalisation by comparing the feature extraction methods with manual gold standard. The results of the evaluation are encouraging: for every news value the majority of features significantly differentiates between the manual annotation labels. This means that for each of the six news values, our approach successfully identified and quantified at least some aspects of that news value.

The evaluation also indicated the need for incorporating world knowledge when analysing headlines. We reliably extract the news value of Proximity, however our current approach relies on a manually created wordlist of UK/US-related terms. A more generic approach for implementing Proximity would require world knowledge to be able to detect that an entity is related to the location of the reader. This would also apply to the implementation of Prominence, whereby entities identified in headline text can be related to the country of the reader. We will address this in Chapter 9.

Our main contributions for this chapter are:

(i) operationalising Magnitude, Proximity, Surprise, and Uniqueness, which have not been used to model social media popularity of news articles before

(ii) implementing novel methods including: using wikification and burstiness for Prominence, and using connotations and bias as indirect measures of Sentiment.

News values which we operationalised in this chapter provide a journalistic perspective for modelling social media popularity of news articles. Using news values we will be able to suggest to a headline author what aspects of the main content to stress (e.g including entities to raise Prominence or making sure that the headline is dissimilar to what has been published recently). However, we also need to know how the phrasing of a headline can be altered to reach higher popularity on social media. To do that we investigate the linguistic style of headlines. In the next chapter we present the implementation details for linguistic style features.

Chapter 5

Implementing Linguistic Style

In order to model the social media popularity of news articles we extract two types of features from headline text. News values (presented in Chapter 4) offer a journalistic perspective on how certain aspects of the news content can be highlighted in the headline for higher social media popularity. Linguistic style features (which we present in this chapter) offer insights into how a headline can be formulated, so that it is more likely to be shared on social media. Compared to news values, headline style is easier to edit, which would be an important factor in applications such as creative writing support software.

To identify relevant linguistic style features for headlines we investigated the literature on news writing and news outlet style guides on headlines (cf. Section 2.5.3). We categorised them into seven feature groups: Brevity, Simplicity, Unambiguity, Punctuation, Nouns, Verbs, and Adverbs. Our implementation of these features is informed by the research on the effect of wording on online popularity of short texts (cf. Section 2.4).

Unlike the previous chapter where we outlined a novel set of features (which required a thorough evaluation), the concepts in this chapter are already clearly defined and have been used in the literature before.

5.1 Implementation

In this section we present the implementation of linguistic style features that we selected (cf. Section 2.5.3). The implementations are summarised in Table 5.1.

Table 5.1: Summary of style feature implementations categorised by feature group. Notation is explained in Section 3.1.2.

Feature name	**Implementation**
BREVITY	
number of words	$\lvert H \rvert$
number of characters	# characters
SIMPLICITY	
parse tree height	parse tree height
number of non-terminal nodes	# non-terminal nodes in parse tree
entropy	entropy of headline
difficult words	$\frac{\text{\# difficult words}}{\lvert H \rvert}$
information content	median information content for nouns and verbs
word frequency	median of word frequencies for all content words
UNAMBIGUITY	
modality	1 if modal event in H or modal relation in H, else 0
number of senses	median number of senses for all content words
PUNCTUATION	
exclamation mark	1 if exclamation mark in H, else 0
question mark	1 if question mark in H, else 0
quote marks	1 if single/double quote marks in H, else 0
NOUNS	
headlinese	1 if # consecutive common nouns $\geq$ 3, else 0
proportion of noun phrases	$\frac{\text{\# NPs}}{\text{\# phrases}}$
proportion of nouns	$\frac{\text{\# common nouns}}{\lvert H \rvert}$
proportion of proper nouns	$\frac{\text{\# proper nouns}}{\text{\# nouns}}$
VERBS	
proportion of verbs	$\frac{\text{\# verbs}}{\lvert H \rvert}$
proportion of verb phrases	$\frac{\text{\# VPs}}{\text{\# phrases}}$
ADVERBS	
proportion of adverbs	$\frac{\text{\# adverbs}}{\lvert H \rvert}$

5.1.1 Brevity

Our implementation of Brevity has two features: the **number of words** and the **number of characters** in a headline. Measuring the text length using both words and characters follows the approach in Arapakis et al. (2014).

5.1.2 Simplicity

We explore two aspects of simplicity: syntactic and lexical. We measure syntactic simplicity with **parse tree height** and the **number of non-terminal tree nodes** in the parse tree. This follows literature on readability, for example Feng et al. (2010).

Lexical simplicity is implemented using four features. The first two – a headline's **entropy** and proportion of **difficult words** – are obtained from a trigram language model. The language model was built using the CMU-Cambridge Toolkit (Clarkson and Rosenfeld, 1997) on the *New York Times* section of the Gigaword corpus (Graff et al., 2003). We define a difficult word as any word not occurring among the 5000 most common words in the language model. The third lexical simplicity feature is median **word frequency** for the content words in the headline. We obtain word frequencies using unlemmatised lists from Word Frequency Data[1] – British National Corpus for *The Guardian* dataset, and Corpus of Contemporary American English for *New York Times* dataset[2]. The fourth feature is median **information content** calculated for nouns and verbs on British National Corpus using NLTK (Bird et al., 2009).

5.1.3 Unambiguity

Our implementation focuses on linguistic expressions of ambiguity, especially through lexical or syntactic means. To capture lexical unambiguity, we calculate the median **number of senses** per word using WordNet – more senses per word indicate a higher chance for ambiguity.

For syntactic unambiguity, we look at **modality**, which expresses unambiguity using syntactic means through the use of modal verbs or modal relations between events. For example, "An incident took place" versus "An incident might have taken place". To check for modality, we use the TARSQI toolkit from the TimeML framework[3]. The TARSQI toolkit looks at both lexical and syntactic modality markers. This allows us to implement a binary feature which indicates if there is a modal event (e.g. *should*) or a modal relation between events (e.g. the modal relation between *promises* and *abolish* in Headline 5.1).

[1] http://www.wordfrequency.info/ [Accessed 13th April 2018]

[2] Although we use word frequencies specific for the news domain, Word Frequency Data also provides general word frequencies which can be used in other domains

[3] http://www.timeml.org/tarsqi/toolkit/index.html [Accessed 13th April 2018]

Headline 5.1. "Michael Gove promises to abolish illiteracy and innumeracy in UK"

5.1.4 Punctuation

We implement three binary features to indicate whether any of the three punctuation marks which are mentioned in headline style guides (**exclamation mark**, **question mark**, **quote marks**) are present in the headline.

5.1.5 Nouns

We implement '**headlinese**' as a binary feature which is positive if there are three or more consecutive nouns. We also look at the overall proportion of **common nouns** and proportion of **proper nouns** to all words and the proportion of **noun phrases** to all syntactic phrases.

5.1.6 Verbs

We implement two features: the proportion of **verb phrases** to all syntactic phrases and the proportion of **verbs** to all words.

5.1.7 Adverbs

We use the proportion of **adverbs** as a feature.

5.2 Linguistic Style Features in Headlines Corpora

In this section we describe linguistic style features as applied on our two headlines corpora. We look in particular at the implementation (cf. Table 5.1) and descriptive statistics in *The Guardian* and *New York Times* corpora (cf. Table 5.2). Since both corpora can be characterised as broadsheet newspapers, we only comment about this particular news writing style.

Brevity has a prevalence of 100%, since some content is always required in a headline. *The Guardian* headlines are longer, both in terms of words and characters, compared to *New York Times*. The median number of words in a headline (*The Guardian*: 10, *New York Times*: 8) shows that (at least) online editions of these two news outlets have fairly long headlines.

Simplicity features have a prevalence of 100% in all cases with the exception of the difficult words feature (*The Guardian*: 95%, *New York Times*: 88%) and information content in *The Guardian* (98%). For syntactic simplicity (parse tree height and number of non-terminal nodes) while the median values for the two corpora are quite similar, the maximum values are much higher for *The Guardian* corpus. A qualitative analysis of some

Table 5.2: Style feature statistics on *The Guardian* and *New York Times* corpora. Reported measures: median and maximum values, prevalence (proportion of non-zero scores).

Feature name	*The Guardian*			*New York Times*		
	Med.	Max.	Prev.	Med.	Max.	Prev.
number of words	10	29	100%	8	18	100%
number of characters	62	160	100%	45	129	100%
parse tree height	8	25	100%	7	15	100%
number of non-terminal nodes	10	34	100%	7	20	100%
entropy	11.29	23.68	100%	11.40	23.68	100%
difficult words	0.30	1	95%	0.38	3	88%
information content	7939	3835155	98%	5704	5799472	100%
word frequency	60.86	1884.27	100%	45.91	1490	100%
modality	0	1	9%	0	1	0%
number of senses	4	36	100%	3.50	36	100%
exclamation mark	0	1	1%	0	1	0%
question mark	0	1	7%	0	1	2%
quote marks	0	1	8%	0	1	1%
headlinese	0	1	27%	0	1	37%
proportion of noun phrases	0.60	1	100%	0.71	1	100%
proportion of nouns	0.27	1	96%	0.08	1	53%
proportion of proper nouns	0.60	1	96%	0.14	1	53%
proportion of verb phrases	0.18	0.86	72%	0	1	39%
proportion of verbs	0.11	0.67	75%	0	0.67	46%
proportion of adverbs	0	0.50	16%	0	1	9%

of *The Guardian* headlines with high syntactic simplicity scores (i.e. with complex syntax) reveals that they often belong in the Opinion genre. The headlines of Opinion pieces in *The Guardian* are often full complex sentences (e.g. Headlines 5.2 and 5.3) which result in larger parse trees. In terms of lexical simplicity, *The Guardian* and *New York Times* have similar median and maximum values for entropy and proportion of difficult words, however in case of information content and average word frequency, *The Guardian* has higher values. In our dataset the news writing in *New York Times* headlines is more complicated than the one in *The Guardian* headlines, evidenced by the lower values for word frequency and slightly higher proportions of difficult words in *New York Times* headlines.

Headline 5.2. "For Australia to deport refugees to Cambodia would be absurd"

Headline 5.3. "The more intelligence I read, the more conservative I become"

Unambiguity reveals few differences between the two news outlets. The statistics for lexical unambiguity (number of senses) show almost no differences. As for modality (our measure of syntactic unambiguity), it is very rare in the corpora (*The Guardian*: 9%, *New York Times*: 0%). All in all, the results on our datasets show that for these two news outlets ambiguous news writing in the headlines is rare.

Punctuation features are among the least prevalent – from 0-1% for exclamation marks, 2-7% for question marks, to 1-8% for quote marks. The median values of 0 mirror these prevalence results. This low prevalence can be explained by the explicit instruction in the style guides to avoid these punctuation marks.

Nouns are more common in *The Guardian* than in *New York Times* (prevalence for the three features which calculate proportions of different types of nouns in *The Guardian* is at least 96%, compared to 53% in *New York Times*). The median values for the proportions of common nouns and proper nouns are also considerably higher in *The Guardian* compared to *New York Times*. This is surprising, as the prevalence of 'headlinese' (i.e. noun clusters) is lower in *The Guardian*.

Verbs are more prevalent in *The Guardian* (72-75% compared to 39-46% in *New York Times*). The median values are also slightly higher in *The Guardian*, although in both cases the values are lower than those for nouns. This finding can be linked to the more complex syntax used in *The Guardian*, as longer and complex sentences require the use of verbs.

Adverbs are quite uncommon (*The Guardian*: 16%, *New York Times*: 9% prevalence) and the median values are 0 in both cases.

In comparison with news values (cf. 4.2 in Chapter 4) the prevalence of style features is much higher. This is to be expected, due to the differing natures of news values and linguistic style features. Style features refer to standard news writing practices which are present in all (or most) headlines. On the other hand, news values are optional. For example, while news values like Prominence or Magnitude would not necessarily be used in some genres like recipes, those headlines still abide by style guidelines. This high prevalence of style features will be useful for building prediction models (cf. Chapter 8), since style features which correlate highly with social media popularity are likely to be found in most headlines.

5.3 Summary

Our goal is to model social media popularity of news articles using news values and linguistic style features in headlines. In this chapter we described the implementation of linguistic style features in headline text. The linguistic style aspects we implemented were: Brevity, Simplicity, Unambiguity, Punctuation, Nouns, Verbs, and Adverbs. We utilised a number of NLP methods, including parsing, language modelling, and event extraction. We applied these methods on two broadsheet headlines corpora and calculated descriptive statistics. We found that most style features are present (i.e. have non-zero scores) in majority of headlines, which can be beneficial when building the prediction models in Chapter 8.

The linguistic style features presented in this chapter are crucial for understanding the effect of headline *wording* on social media popularity. By implementing a wide variety of linguistic style features we will be able to understand how a headline can be reformulated to reach higher popularity on social media.

Our main contribution for this chapter is:

(i) implementing seven aspects of linguistic style for extraction from headline text

We would like to note that, as with any NLP task, the reliability of the feature engineering rests on the accuracy of the text processing tools. Some features are anticipated to have very small error rates (Brevity, Punctuation), however parsing might be more prone to errors due to the limited context. Although we conducted sanity checks and manually checked small samples of processed text, we acknowledge that any mistakes in the initial processing (e.g. part-of-speech tagging or parsing) will have impact on any downstream applications.

The implementation of linguistic style features completes the third step in our experimental methodology (cf. Section 1.4). In the next two chapters we turn to the fourth step in our methodology – investigating the impact of news values and linguistic style features. In the next chapter we correlate the features we extracted from headline text (news values and linguistic style) with social media popularity. By doing that we are able to investigate which headline features should be maximised in order to achieve higher social media popularity. This is a crucial insight for any headline author.

Chapter 6

Impact of News Values and Style on Social Media Popularity

In order to model the social media popularity of news articles using headlines we collected and preprocessed news article headlines from two broadsheet news sources (*The Guardian* and *New York Times*). We also obtained the social media popularity on Twitter and Facebook for each news article. Our hypothesis is that the way we formulate headlines influences social media popularity of news articles. To investigate that we proposed using two feature groups: news values which offer a journalistic perspective, and linguistic style which offers insights into wording. We described our methods for implementing news values in Chapter 4 and for linguistic style in Chapter 5.

In this chapter we investigate whether news values and style features correlate significantly with social media popularity. By that we mean whether certain characteristics of headline text are associated with higher or lower levels of some forms of engagement on social media: tweeting, liking, sharing. This investigation is crucial for identifying headline aspects which can be added, highlighted, or changed, so that a rephrasing of the headline results in higher social media popularity. We do this by correlating feature values for news values and linguistic style with social media popularity metrics. Details on the social media popularity metrics we used and their data collection method are described in Section 3.2.

6.1 Method

Our goal is to correlate the values for each news value and linguistic style feature with social media popularity. The correlations are calculated on the training sets separately for each news source. In our datasets we have two types of features: binary and numeric. For numeric features we use the Kendall rank correlation coefficent (Kendall's τ, Kendall (1938)). This correlation coefficient was chosen because it can be used with non-parametric data and is suited for datasets with tied ranks. For binary features we check whether the feature median is significantly higher or lower than the overall median for that social media popularity measure. We compare the medians using the Wilcoxon signed rank test (Wilcoxon, 1945).

6.2 Results

Results of these calculations are reported in Table 6.1 for *The Guardian* and Table 6.2 for *New York Times*. The correlations will be discussed in Section 6.3 for news values and in Section 6.4 for linguistic style. In order to gain deeper insights into the behaviour of features we compare the results on the two corpora and try to identify similarities and differences.

Since news values were originally formulated for typical news articles (e.g. about news events) we consider separately correlations for the news subset (News in Tables 6.1 and 6.2) and articles across all genres (All in Tables 6.1 and 6.2). The method for obtaining the news subsets is described in Section 3.1.3.

Table 6.1: Feature correlations with social media popularity for *The Guardian* dataset. News = news subset, All = all genres. Numeric features: Kendall's τ (* p<0.05, ** p<0.01); binary features: feature median is higher/lower (↑ / ↓ p<0.05) than the overall median (in brackets in column headings) calculated using Wilcoxon signed rank test.

		News		**All**	
		(3,161 articles)		**(11,980 articles)**	
	Feature name	T (57)	F (83)	T (41)	F (42)
NEWS VALUES FEATURES					
Prominence	number of entities	0.04**	0.01	0.07**	0.03**
	News recent prominence	0.05**	0.01	0.11**	0.02**
	Wikipedia long-term prominence	0.08**	0.05**	0.16**	0.11**

table continues

continue table

		News		All	
	Feature name	T (57)	F (83)	T (41)	F (42)
	Wikipedia day-ago prominence	0.08**	0.05**	0.15**	0.11**
	Wikipedia current burst size	0	-0.01	0	0.01
	Wikipedia burstiness	0.05**	0.02	0.07**	0.02**
Sentiment	sentiment	-0.02	0.01	-0.06**	-0.04**
	polarity	0.09**	0.06**	0.1**	0.09**
	connotations	0.06**	0.04**	0.05**	0.06**
	bias	0.05**	0.03*	0.07**	0.06**
Magnitude	comparative/superlative	0.06**	0.03*	0.03**	0.03**
	intensifiers	0.06**	0.05**	0.04**	0.03**
	downtoners	0.05**	0.07**	0.03**	0.02**
Proximity	proximity	35↓	38↓	40	34↓
Surprise	surprise	-0.04**	-0.03*	-0.02**	-0.01
Uniqueness	headline uniqueness	-0.06**	-0.04**	-0.06**	-0.08**
STYLE FEATURES					
Brevity	number of words	0.13**	0.12**	0.14**	0.11**
	number of characters	0.09**	0.07**	0.13**	0.09**
Simplicity	parse tree height	0.09**	0.08**	0.15**	0.12**
	number of tree nodes	0.11**	0.1**	0.15**	0.12**
	entropy	-0.08**	-0.1**	-0.07**	-0.1**
	proportion of difficult words	-0.05**	-0.04**	-0.06**	-0.06**
	information content	0.03*	0.02*	0.09**	0.07**
	word frequency	-0.05**	-0.03**	0	-0.03**
Unambiguity	number of senses	0.03*	0.03*	0.01*	0
	modality	43↓	39↓	57↑	64↑

table continues

continue table

		News		**All**	
	Feature name	T (57)	F (83)	T (41)	F (42)
Punctuation	exclamation mark	35	21↓	20↓	33
	question mark	41↓	52↓	50↑	69↑
	quote marks	37↓	43↓	49↑	68↑
Nouns	three consecutive nouns	38↓	42↓	40	35↓
	NP count	-0.11**	-0.09**	-0.14**	-0.1**
	proportion of nouns	-0.02	-0.06**	0.03**	0.03**
	proportion of proper nouns	0.07**	0.06**	0.09**	0.1**
Verbs	VP count	0.11**	0.08**	0.13**	0.09**
	proportion of verbs	0.08**	0.06**	0.13**	0.1**
Adverbs	proportion of adverbs	0.09**	0.11**	0.04**	0.04**

Table 6.2: Feature correlations with social media popularity for *New York Times* dataset. News = news subset, All = all genres. Numeric features: Kendall's τ (* p<0.05, ** p<0.01); binary features: feature median is higher/lower (↑ / ↓ p<0.05) than the overall median (in brackets in column headings) calculated using Wilcoxon signed rank test.

		News		**All**	
		(3,843 articles)		**(5,074 articles)**	
	Feature name	T (117)	F (200)	T (102)	F (153)
News Values Features					
Prominence	number of entities	-0.06**	-0.06**	-0.02	-0.01
	News recent prominence	0.11**	0.07**	0.11**	0.06**
	Wikipedia long-term prominence	0.1**	0.05**	0.11**	0.08**
	Wikipedia day-ago prominence	0.12**	0.07**	0.13**	0.09**
	Wikipedia current burst size	0.06**	0.06**	0.08**	0.08**
	Wikipedia burstiness	0.04**	0.01	0.06**	0.03**

table continues

continue table

		News		**All**	
	Feature name	T (117)	F (200)	T (102)	F (153)
Sentiment	sentiment	0	0.02	0.01	0.02
	polarity	0.09**	0.1**	0.11**	0.12**
	connotations	0.01	0.02	0.05**	0.06**
	bias	0.08**	0.07**	0.09**	0.08**
Magnitude	comparative/superlative	0.02	0.02	0.04**	0.03**
	intensifiers	0.01	0.02	0.03**	0.04**
	downtoners	0.03*	0.01	0.04**	0.03*
Proximity	proximity	39↓	38.5↓	40↓	40↓
Surprise	surprise	-0.02	-0.02	-0.01	-0.01
Uniqueness	headline uniqueness	0.02	-0.01	0.01	-0.02
STYLE FEATURES					
Brevity	number of words	0.16**	0.1**	0.2**	0.15**
	number of characters	0.15**	0.08**	0.19**	0.12**
Simplicity	parse tree height	0.13**	0.11**	0.16**	0.13**
	number of tree nodes	0.15**	0.11**	0.17**	0.15**
	entropy	0	-0.07**	0.03**	-0.04**
	proportion of difficult words	-0.07**	-0.03**	-0.06**	-0.02
	information content	0.05**	0.04**	0.02*	0
	word frequency	0	-0.02*	0.01	-0.02*
Unambiguity	number of senses	0.01	0	-0.01	-0.01
	modality	30↓	25↓	46↓	42↓
Punctuation	exclamation mark	26↓	9↓	31↓	63.5
	question mark	39↓	35↓	37↓	58↓
	quote marks	41↓	32↓	41↓	35↓

table continues

continue table

		News		**All**	
	Feature name	T (117)	F (200)	T (102)	F (153)
Nouns	three consecutive nouns	38↓	38↓	40↓	42↓
	NP count	-0.11**	-0.13**	-0.12**	-0.14**
	proportion of nouns	0.05**	0.02	0.04**	0.03**
	proportion of proper nouns	0.06**	0.03**	0.05**	0.04**
Verbs	VP count	0.07**	0.1**	0.1**	0.12**
	proportion of verbs	0.08**	0.09**	0.11**	0.11**
Adverbs	proportion of adverbs	0.06**	0.06**	0.06**	0.07**

6.3 Impact of News Values

We first look at the impact of six news values on news article popularity on Twitter and Facebook. We discuss each news value and its impact in the following sections.

6.3.1 Prominence

Most features which implement this news value correlate significantly with social media popularity. This is shown for both *The Guardian* and *New York Times* headlines.

Number of entities correlates positively with most measures in *The Guardian*, but the correlation is negative for the news subset in *New York Times*. This means that while including more entities is linked with higher news article popularity for *The Guardian* headlines, it is the opposite case for *New York Times*. This might be due to capitalisation of *New York Times* headlines which could have introduced noise at the preprocessing stage which included named entity recognition.

Recent prominence in news source headlines (news recent prominence feature) is significantly positively correlated for all metrics with the exception of news subset in *The Guardian*. The effect is strongest for Twitter measures across all genres (τ = 0.11).

The two prominence features which use Wikipedia as prominence source achieve some of the strongest correlations among all news values features (up to τ = 0.16). Both features achieve a significant positive correlation with all measures, with slightly higher correlations for Twitter measures. This finding supports one of the key journalistic hypotheses about news values – that prominence of entities is an important factor for news article popularity.

The ‘bursty’ aspect of entities’ prominence which we explore for the first time for modelling news article popularity is captured with two features. For both features at least half of the measures correlate significantly with social media popularity. Firstly, for current burst size the correlations are statistically significant only for *New York Times* measures and the correlation is positive, however none of *The Guardian* measures reached p<0.05 significance level. We suggest that this finding is due to differing preferences of audiences of *The Guardian* and *New York Times*, since the prevalence of this feature is comparable between the two news sources (*The Guardian*: 12%, *New York Times*: 10%; cf. Table 4.2). It might also indicate that although an entity is in a burst when considering global prominence on Wikipedia, this burst does not necessarily carry over to the national level. Secondly, burstiness (which considers number of days that an entity has been in a burst over a year) is significantly positively correlated for all measures with the exception of Facebook in the news subset. Significant correlations with at least half of the popularity measures for these two features confirms our hypothesis that temporal aspects of prominence (‘burstiness’) influence popularity of news articles to a certain extent.

Overall, most Prominence features correlate significantly with social media popularity. The Facebook popularity measure for *The Guardian* news subset has fewest significant correlations – this might be due to differences in sharing behaviour across news sources and social media platforms. Prominence is the news value that has been mentioned the most in journalism studies literature and independently has been implemented in news article popularity prediction tasks. We add to that work on Prominence by: (i) introducing wikification and burstiness to the implementation, (ii) extracting Prominence from headlines, and (iii) carrying out an in-depth analysis of its correlations with Twitter and Facebook popularity using two corpora. Our findings make a strong case for the importance of mentioning prominent entities in headlines.

6.3.2 Sentiment

In total we implemented four Sentiment features. The first two (sentiment and polarity) considered the direct sentiment value of the headline. Sentiment (whether the headline was more positive or negative; the lower the value, the more negative the headline) only has a significant effect on *The Guardian* measures in the all genres setting. The correlation is negative, which means that more negative headlines attract more attention. This follows previous research (e.g. Galtung and Ruge (1965) in journalism studies and Reis et al. (2015) in computer science). On the other hand, polarity (which considers the total value of sentiment) correlates positively with all measures. In case of all genres setting in *New York Times* the correlations are some of the highest for news values. This finding shows that regardless of whether they are positive or negative, sentiment-charged headlines tend to be more popular on social media.

The other two Sentiment features (connotations and bias) target indirect sentiment. With the exception of *New York Times* news subset, connotations correlate positively with social media popularity. In case of the bias feature, correlations are significantly positive for all measures. Taken together, these results point to indirect sentiment also having an impact on popularity. Our decision to broaden the scope of Sentiment implementation to take into account indirect sentiment (through connotations and biased language) is proven correct, since these features are shown to be significantly correlated with social media popularity.

Overall, sentiment-charged and biased language in headlines are positively correlated with social media popularity. Our findings show that any sentiment-charged language in headlines – regardless of whether the sentiment is negative or positive – has a significant positive correlation with social media popularity, which is in line with findings by Reis et al. (2015). We add to previous research by showing that indirect sentiment (bias, connotations) also significantly correlates with social media popularity.

6.3.3 Magnitude

We implemented three features for this news value. The first feature is based on part-of-speech tags which indicate magnitude (comparative and superlative adjectives and adverbs). The feature is significantly positively correlated for all measures with the exception of the *New York Times* news subset. The highest correlation (τ = 0.06) for that feature is with Twitter popularity in *The Guardian* news subset.

The other two Magnitude features we implement use particular lexical categories which enhance (intensifiers) or diminish (downtoners) the words they describe. According to the journalistic perspective on Magnitude intensifiers should be positively correlated with popularity, while downtoners should be negatively correlated. We found that in both cases the correlations are significantly positive (with the exception of *New York Times* news subset where the correlations are not significant). This suggests that any enriching of descriptors in the headline – be they intensifiers or downtoners – correlates positively with popularity on social media.

Overall, Magnitude features are significantly positively correlated with the majority of social media popularity measures. Nearly all measures in *New York Times* news subset are not significant. On the other hand, *The Guardian* news subset has the highest correlations for this news value (up to τ=0.07). It is interesting to note that relatively infrequent lexical items such as intensifiers and downtoners (6-10% prevalence in our headlines corpora for intensifiers and 3-4% prevalence for downtoners) correlate significantly with the popularity of headlines. In particular, it is somewhat surprising that the presence of downtoners, which function to diminish the lexical items they describe, also correlates positively with popularity. This might be due to a downtoner being used as an oxymoron which actually intensifies the scope through a rhetorical device (e.g. Headlines 6.1 and 6.2). We are the first to investigate whether Magnitude in headlines correlates with social media popularity.

Headline 6.1. "Why the Almeida is a little wonder"

Headline 6.2. "Plucky Little Manchester United"

6.3.4 Proximity

Our first implementation of Proximity for the task of modelling social media popularity focused on geographic proximity. We assumed that readers from the same country as the news outlet constitute a large part of its readership, and we looked only at geographic proximity to the news source. That is to say, for *The Guardian* we checked whether there were any UK-related mentions in the headline, and for *New York Times* we checked for US-related mentions. We found that there is a statistically significant effect for nearly all social media popularity measures, however that effect is negative. This means that if there is a UK-related entity in a headline from *The Guardian* that headline has a lower than average popularity (median Facebook popularity for a news article *The Guardian* is 42; with a UK-related entity in headline it is 34). It is the same case for *New York Times*, however the differences are larger (median Facebook popularity for a news article in *New York Times* is 153; with a US-related entity in the headline it is 40).

Overall, the Proximity feature in headlines has a significant – but contrary to our expectations – negative effect on social media popularity. This finding (which has not been investigated for headlines before) suggests that either: (i) *The Guardian* and *New York Times* are more oriented towards international news, or (ii) the reader location needs to be controlled in order to get a more accurate reflection of *The Guardian* and *New York Times* readership. We address the latter point in Chapter 9, where we take into account readers' country location when modelling social media popularity.

6.3.5 Surprise

We implemented one Surprise feature. It considered the commonness of syntactic chunks in the headline – the more uncommon the phrase, the more surprising it is. We found that this feature only achieves a statistically significant correlation with three out of eight social media popularity measures. The three significant correlations (all for *The Guardian*) are negative. This follows our expectations, since the lower the value of this feature (i.e. lower log-likelihood), the more surprising the phrasing of the headline is. Louis and Nenkova (2013), who used a similar approach to identify 'creative language' in science articles, also found that uncommon word combinations positively influenced popularity.

Overall, although this news value only achieves significant correlations with just under half of the social media popularity measures we used, the direction of the significant correlation follows the literature on news values and findings for related tasks. We add to that research by investigating surprising phrasing in headlines.

6.3.6 Uniqueness

Our implementation of the news value of Uniqueness calculated similarities between a given headline and recently published headlines. The feature has a significant correlation with all popularity measures in *The Guardian* dataset, but not in *New York Times*. This might be specific to the news outlet. For example, *New York Times* tends to publish similar headlines as regular features (cf. Headlines 6.3-6.5), which might not negatively influence readers when sharing news articles on social media. On the other hand, *The Guardian* has many examples of repeating a headline when they publish a related article which contains audio or video (cf. Headlines 6.6-6.7 and Headlines 6.8-6.10), where the latter articles are not usually shared on social media. This is corroborated for *The Guardian* where the correlation is negative, which means that if there is a very similar headline in recent past, then the current headline will tend to be less popular.

- Examples of lack of Uniqueness in *New York Times*:

  Headline 6.3. "36 Hours in [place name]"

  Headline 6.4. "Comedy Listings for [date]"

  Headline 6.5. "Fantasy Football: Week [number]"

- Examples of lack of Uniqueness in *The Guardian*:

  Headline 6.6. "A day in the life of a journeyman jockey"

  Headline 6.7. "A day in the life of a journeyman jockey – in pictures"

  Headline 6.8. "Hong Kong pro-democracy march attracts tens of thousands"

  Headline 6.9. "Hong Kong pro-democracy march – timelapse video"

Headline 6.10. "Hong Kong pro-democracy protest – in pictures"

Overall, the Uniqueness feature is significantly negatively correlated with social media popularity, but only for *The Guardian* datasets. We are the first to investigate whether Uniqueness in headlines correlates with social media popularity.

6.4 Impact of Linguistic Style

We now turn to linguistic style features. The correlations of style features with social media popularity for each style aspect are discussed below.

6.4.1 Brevity

We implemented Brevity as headline length in words and in characters. Both features are significantly positively correlated with social media popularity for all measures. Indeed, when looking at all genres in *New York Times*, the correlations with Twitter popularity are the strongest out of all features (up to τ = 0.2). A positive correlation means that longer headlines tend to be more popular on social media. This of course goes against the traditional journalistic requirement for headlines to be brief. However, in the context of online news consumption where many readers do not click past the headline (cf. Section 2.3), longer – and thus more informative – headlines are more popular.

Overall, Brevity features achieve some of the strongest correlations for our datasets. The positive correlation goes against traditional journalistic expectations, but might be explained by the current online news reading behaviours. Arapakis et al. (2017) also reported significant positive correlations with social media popularity for number of words and number of characters in news articles.

6.4.2 Simplicity

We implemented six features to capture Simplicity. The first two – parse tree height and number of parse tree nodes – explore syntactic simplicity. For both features there are significant positive correlations with all social media popularity measures. They are particularly strong for Twitter popularity in both *The Guardian* and *New York Times* all articles setting. The positive correlation is surprising – it means that the more complex the syntactic structure, the more popular the headline tends to be on social media. It probably links with the informativeness aspect which is important for 'headline gazers' (cf. Section 2.3), as syntactically complex headlines might convey more information.

The next four features consider lexical simplicity. Entropy is significantly negatively correlated for nearly all measures. The strongest correlations are for Facebook measures in *The Guardian* (τ = -0.1). Proportion of difficult words is also significantly negatively correlated for nearly all measures. In this case the strongest correlation is with Twitter popularity in *New York Times* news subset. Information content has a significant positive impact for nearly all popularity measures. The strongest correlation is with Twitter popularity for *The Guardian* dataset. Finally, word frequency is significantly negatively correlated for five out of eight social media popularity measures, including all Facebook popularity measures. These correlations are some of the lowest among the Simplicity features reaching τ = -0.05 for Twitter popularity in *The Guardian* news subset. In terms of the expected direction of the correlation, the first three lexical features (entropy, proportion of difficult words, information content) follow expectations in that simpler vocabulary is likely to occur in more popular headlines. The last lexical feature (word frequency) has a negative correlation with popularity, meaning that the less frequent the words (and thus potentially more complex) are in the headline, the higher the popularity of the news article. The relatively sophisticated language in both *The Guardian* and *New York Times* (they are both broadsheet newspapers), as well as the considerable presence of topics such as business or technology (cf. Section 3.1.3 which includes category statistics) could explain this result.

Overall, we are the first to find that most Simplicity features in headlines have a significant correlation with social media popularity, including some of the highest correlations in this dataset in the case of syntactic simplicity features. However, we found the direction of some of the significant correlations (namely syntactic simplicity and word frequency) to be the opposite to our expectations or the expectations from literature, which called for simplicity in all cases. We hypothesise that these unexpected findings can be explained by the need for informativeness and the relatively sophisticated language used in broadsheet news outlets.

6.4.3 Unambiguity

We implement two Unambiguity features. The first one (number of senses) considers potential lexical ambiguity. The feature significantly correlates with only three out of eight measures (all in *The Guardian* dataset) and the correlation is very low (up to τ=0.03). The positive direction of the correlation is also surprising. The other feature (modality) uses syntactic information to consider event or author ambiguity (e.g. something *might* happen). The feature shows a significant correlation with all social media popularity measures. Since presence of modality could lead to ambiguity, we expected a negative impact on social media popularity. That is indeed the case for most measures. However, for all articles setting in *The Guardian* the effect is positive – meaning that headlines with some modality element tend to be more popular on social media.

Overall, the results for the Unambiguity features are inconclusive. For one feature (number of senses) we have only several significant correlations and the direction of the correlation is positive, whereas we expected it to be negative (i.e. less senses = less ambiguity = more popular). For the other feature (modality) we have a significant association with all social media popularity measures, but for one dataset (*The Guardian* all genres) the effect is the opposite to what we expected. Modality is also one of the less prevalent style features, so we cannot draw firm conclusions. Unambiguity in headlines has not been investigated before in relation to social media popularity.

6.4.4 Punctuation

We implement three binary features to check for the presence of a given punctuation mark in the headline. Following the advice from the style guides we expect the association to be negative. For all three features, there is a significant correlation with most popularity measures (there are some measures for which the effect is not significant in case of exclamation marks). As with the modality feature in the previous section, we found that while for the majority of cases we get the expected negative association, for *The Guardian* all articles case we get the opposite (i.e. positive) association for the presence of question marks and question marks.

Overall, we found that certain punctuation marks tend to occur in headlines with lower social media popularity. However, in case of question marks and quote marks, *The Guardian* all articles dataset shows a significant positive association. We would like to note that although we found punctuation to correlate significantly with social media popularity, all punctuation features are quite rare (up to 8% prevalence in *The Guardian* headlines and only up to 2% prevalence in *New York Times* headlines). For a sample of Dutch headlines rewritten for the Blendle news aggregator, Kuiken et al. (2017) reported that using question marks and quote marks had a statistically significant negative effect on headline click-through rate, which aligns with most of our findings.

6.4.5 Nouns

We implemented four features. The first one looks specifically at ‘headlinese’, i.e. whether there are three or more successive nouns in headline. With the exception of Twitter popularity measure in *The Guardian* all articles setting where the result was not significant, ‘headlinese’ tends to occur in headlines with lower social media popularity. The next feature is noun phrases count. For all social media popularity measures there is a significant negative correlation reaching τ = -0.14 for Twitter in *The Guardian* all genres setting and for Facebook in *New York Times* all genres setting. The results for ‘headlinese’ and noun phrase count follow the expectations from literature and news outlet style guides. However, the last two features (proportion of common nouns and proportion of proper nouns) have correlated positively with most social media popularity measures, which is contrary to expectations. As proper nouns indicate a named entity (and we found entities' Prominence to be significantly positively correlated with social media popularity), this can explain the positive correlation for the proportion of proper nouns.

Overall, we found that the presence of some types of nouns has a negative correlation with headline popularity – namely 'headlinese' and noun phrase count (which aligns with findings by Arapakis et al. (2017) that the proportion of nouns in news articles correlates negatively with social media popularity). On the other hand, proportion of proper nouns (which can be linked to Prominence) and in some cases proportion of common nouns shows a positive correlation.

6.4.6 Verbs

We implement two features: the number of verb phrases and the proportion of verbs. Both are positively correlated with all social media popularity measures, reaching τ = 0.13 for Twitter popularity measure in *The Guardian*. This result clearly supports the guidelines included in the style guides, which encourage the use of verbs.

Overall, using more verbs in headlines is significantly positively correlated with social media popularity, which was reported for whole news articles by Arapakis et al. (2017).

6.4.7 Adverbs

We implemented one feature – the proportion of adverbs. We found that there is a statistically significant positive correlation for all popularity measures. The highest correlation (τ = 0.11) was with Facebook popularity in *The Guardian* news subset.

Overall, we found that using adverbs in headlines has a significant positive correlation with social media popularity. When looking at whole news articles, Arapakis et al. (2017) reported more mixed results, namely that the proportion of adverbs was negatively correlated with some social media popularity metrics one hour after publication, but was positively correlated one week after publication.

6.5 Discussion

Overall, we found that each of the features we developed is correlated at a statistically significant level with at least three out of eight social media popularity measures. In many cases there are statistically significant correlations with over half of the popularity measures. This shows there are indicators of social media popularity in headline text. We now discuss several issues which have emerged from our findings.

News values. We implemented news values for the first time for the task of modelling news article popularity on social media. We found that for every news value there were features significantly correlated with social media popularity, which supports our proposal to include news values when researching text content of headlines (and potentially other online artefacts). The news values of Prominence and Sentiment had particularly strong correlations. These types of features have also been used previously for modelling popularity of online content (e.g. Tan et al. (2014); Arapakis et al. (2014)), however we have considerably broadened the scope of these features by utilising wikification and burstiness for Prominence, and considering connotations and biased language as indirect measures of Sentiment. In particular the use of Wikipedia as Prominence source resulted in some of the highest correlations for the datasets. The other news values we implemented have not been considered before for the task of news article popularity prediction. The correlations we report here for these new features add new insights about the formulation of headlines for higher social media popularity. We found that explicit linguistic indicators of Magnitude correlated significantly with social media popularity, although their prevalence in both headlines corpora was quite low (up to 10%; cf. Table 4.2). This suggests that more emphatic language should be used in headlines. Contrary to our expectations for the news value of Proximity, we found that mentioning keywords referencing UK or US was associated with *lower* social media popularity. As this finding could have been influenced by the global nature of the readership of both *The Guardian* and *New York Times*, we conduct additional analysis in Chapter 9 where we control for user location. By comparing the two corpora we found that the impact of some news values can differ depending on the news outlet (e.g. Surprise and Uniqueness where the correlations were significant only for *The Guardian*), while some show significant correlations for all datasets (Prominence, Sentiment). We showed that for majority of cases our implementation of news values adds new insights on formulating headlines to increase social media popularity.

Linguistic style. Our results show that most features relating to the phrasing of headlines were significantly correlated with the social media popularity of news articles. Indeed, the strongest correlations we measured were for style features, such as the number of words or the number of parse tree nodes. This is an important insight, since headline features relating to linguistic style can be easily edited, which can lead to the headline reaching higher popularity on social media. The correlations we report here point to *how* a headline can be edited: write longer, sentence-like headlines which include verbs and proper nouns, but not difficult words.

Some of our findings were unexpected. For example, text length and syntactic complexity features positively correlated with popularity. Some of the previous NLP research on readability (Pitler and Nenkova, 2008) found that text length and parse tree height were negatively correlated with readability. We argue that in case of headlines it makes sense that these features would be positively correlated, as longer headlines with full sentences (and thus larger parse trees) would be easier to understand than the usual headline style exemplified by 'headlinese'. Another unexpected finding is the positive effect of quote marks and question marks in *The Guardian* dataset. A qualitative analysis of some of the more popular headlines with these features found that the presence of quote marks can indicate bias (e.g. Headline 6.11), which is consistent with the positive correlation of the bias feature for that dataset. Question marks can indicate event uncertainty (e.g. Headline 6.12). This can be linked with the results for the modality feature in *The Guardian* all articles dataset, where the feature median was also significantly higher. We note that this applies to a small number of headlines, as both modality and punctuation features have a low prevalence in the corpora.

Headline 6.11. "Spanish celebrate 'conquest' of French politics"

Headline 6.12. "Is the 'cost of living crisis' over?"

Differences between Twitter and Facebook. The correlations with social media popularity for Twitter are slightly stronger than for Facebook. When taking into account news sources as well, in *The Guardian* dataset correlations are higher for Twitter compared to Facebook, however in *New York Times* Facebook usually has higher correlations. This variation aligns with reports which describe differences in demographics of news readers on these two websites (Gottfried and Shearer, 2016), which can impact what kind of headlines are preferred.

Another factor that should be considered is the purpose and usage of these two social media websites. Twitter as a microblogging platform is meant for sharing content to a wide audience and most tweets are public. On the other hand, Facebook also allows sharing, but usually the audience is private (the user's Facebook friends). This difference is usage might influence the *type* of news articles that a reader engages with (by sharing or liking), which in turn might affect what kind of headline wording is preferred by readers.

Differences between news sources. When comparing feature correlations between *The Guardian* and *New York Times*, we found that there are some features which play a particularly significant role for only one news outlet. For example, most news values and verb-related features were more strongly correlated for *The Guardian*; whereas for *New*

York Times it was Sentiment, Brevity and syntactic Simplicity features. Differences in headline writing styles between *The Guardian* and *New York Times* might contribute to the differences in feature correlations. For example, the headlines in the Opinion section in *The Guardian* have a rather distinct, conversational style. As they are among the most popular news articles in our dataset that can influence correlations results.

Scope. It is important to note that our results are limited to features extracted directly from headline text. However, there are some factors (outside of headline text) which may influence news article popularity on social media. These factors include visual presentation (e.g. whether the headline was displayed on the top of the page or otherwise made visually striking), and social network (e.g. whether the headline was tweeted by a high-profile celebrity or shared on Facebook by a friend). A hybrid approach which uses both content (headline text) and context (visual and social information) is proposed for future work (cf. Section 10.4). These confounding factors might contribute to the relatively low correlations (the highest correlation was $\tau = 0.2$). We note that the correlations we found are still comparable – or higher – to the correlations reported by Arapakis et al. (2017), however an exact comparison is impossible because they used slightly different metrics. Compared to Arapakis et al. (2017) we investigated a much wider range of text-derived features, which provides insights about how to formulate a headline to achieve higher social media popularity. We also take a more qualitative perspective by investigating impact of features on perceived popularity in Chapter 7.

6.6 Summary

Our main hypothesis for this thesis is that the way we formulate news article headlines influences their popularity on social media. In the previous chapters we proposed the use of news values and linguistic style as the features to explore formulating headlines. This chapter presents one of our key contributions – correlations of headline-derived feature values with news article popularity on Twitter and Facebook. These results show which features significantly correlate with social media popularity, what is the direction of that correlation, and whether the correlation is significant for both news outlets and both types of social media popularity metrics. These findings are crucial for understanding how a headline can be rephrased, so that it achieves higher social media popularity.

For all news values and linguistic style features we found a statistically significant correlation with at least three out of eight social media popularity measures. This shows that the formulation of headlines correlates with social media popularity of news articles. The features that had the strongest correlations were: Prominence (which proves the benefit of our proposed method using wikification and burstiness), Sentiment (including indirect

Sentiment which we introduced), Brevity, Simplicity (which we considered for this task for the first time), and Verbs. Prominence of entities being an important factor for news article popularity and the unexpected results for the Proximity feature encourage us to further explore these news values by adding country relatedness information from Wikidata and building a country-specific prediction model (cf. Chapter 9).

Our main contribution for this chapter is:

(i) an in-depth investigation of correlations of news values and linguistic style features with social media popularity of news articles, including a discussion of similarities and differences between feature types, news outlets, and social media popularity measures

In this chapter we correlated feature values with social media popularity as measured by the amount of attention that a news article gets on Twitter and Facebook. This gives us insights into how news values and linguistic style aspects correlate with readers' social engagement with the content through tweeting, retweeting, liking, or sharing a news article. We also want to understand *why* readers decide to engage with certain headlines and whether news values and linguistic style of headlines play a role in that decision. In the next chapter we describe the results and analysis of a crowdsourced survey which shows us the impact of features on *perceived popularity*.

Chapter 7

Impact of News Values and Style on Perceived Popularity

In order to model social media popularity of news articles using headlines we need to establish which headline features correlate with popularity. We proposed using news values (Chapter 4) and linguistic style (Chapter 5) for this task. In Chapter 6 we showed that news values and linguistic style in headlines significantly correlate with social media attention – the popularity of *The Guardian* and *New York Times* articles on Twitter and Facebook. We did that by correlating feature values of news values and style with social media popularity measures obtained from Twitter and Facebook.

In this chapter we look at the impact of news values and style on *perceived popularity* – whether or not readers think a headline would be clicked on, and whether news values and linguistic style influence their decisions. We do that by conducting a survey on a crowdsourcing platform and conducting a qualitative analysis of survey responses by three experts. Our goal is to establish whether the news values and style features we propose have an effect on direct (i.e. clicking) as well as social (e.g. tweeting or sharing) engagement. This provides complementary insights to the results from the previous chapter, where we used quantitative methods to find significant correlations of our features with engagement on social media.

7.1 Study Design

In order to obtain the perceived popularity measure and ask readers directly about the influence of news values and linguistic style in headlines on popularity, we conducted a crowdsourced survey. This sections details the study design.

7.1.1 Study Objectives

In this chapter our goal is to answer the following questions:

- Do news values and linguistic style in headline influence readers' perception of popularity?
- Does the impact of news values and linguistic style differ between perceived popularity and social media popularity?
- Can qualitative analysis of findings on perceived popularity help to explain the results of correlations with social media popularity?

Crowdsourced survey. In order to answer these questions we decided to conduct a survey. Through a survey we can control the selection and presentation of headlines (thus eliminating the confounding factor of layout and presentation which readers encounter when reading news online), and ask directly about factors which influence participants' decisions. We decided to use a crowdsourcing platform (CrowdFlower), so that we could recruit participants from around the globe, which mirrors the global readership of the news outlets we use in our headlines corpora.

Perceived popularity. Crowdsourcing has been acknowledged to have issues in terms of reliability of annotations (e.g. Maynard and Bontcheva (2016); Samimi et al. (2017)). Lofi et al. (2012) proposed a classification of crowdsourcing tasks. One of the axes of comparison was 'level of answer ambiguity/agreement', where tasks ranged from factual tasks (e.g. "Is there a person in this photo?"), through consensus tasks (e.g. "Does the person on this photo look happy?"), to opinionated tasks (e.g. "What is the nicest colour?"). We argue that our task of inducing judgements about headlines should be seen as opinionated. Although there might be a consensus for some headlines, overall the effects of personal preferences would be too strong. This has been shown to be the case for other research tasks. For example, Figueiredo et al. (2014) used crowdsourcing to investigate whether video content influences popularity on YouTube. They found that participants often could not agree which video would be more popular, because their perception of the content was very subjective. In our survey we follow the study design by Tan et al.

(2014) who asked about the preferences of *other people*, in order to obtain more objective responses. This would still result only in perceived popularity, due to self-enhancement bias (Pronin et al., 2004), meaning that people overestimate the prevalence of personal opinions among other people.

Methods. We make use of both quantitative and qualitative methods in order to gain deeper insights into the effects of news values and linguistic style on perceived popularity. Similarly to the previous chapter we calculate correlations between feature values and perceived popularity measure (cf. Section 7.2). We also conduct a qualitative analysis of survey responses by three experts (cf. Section 7.3) and look at the feedback from CrowdFlower participants on the survey (cf. Section 7.1.4).

7.1.2 Survey Content

We conducted a survey through CrowdFlower, a crowdsourcing platform. The study design was approved by the Mathematics and Physical Sciences and Engineering Joint Faculty Research Ethics Committee at the University of Leeds (application reference: MEEC16-003). The full survey is presented in Appendix E and summarised in this section. The survey was split into four sections:

Preliminaries: information sheet and consent form, questionnaire on demographics and news reading habits

PART 1 Headline popularity:

- 48 headlines (randomly sampled; two versions: *The Guardian* and *New York Times*)
- For each headline participants were asked: *How likely is it that other people will click on this headline?*
- five Likert scale responses (*Extremely likely, Slightly likely, Neutral, Slightly unlikely, Extremely unlikely*)

PART 2 Judgement criteria:

- 12 features of headlines (five news values and seven linguistic style aspects)
- For each feature we presented a short definition and several examples of its usage
- Participants were asked to indicate to what extent each feature influences the decision about clicking on headlines: (i) for them personally (*I personally*

consider this feature when clicking on headlines), and (ii) for other people (*I think this feature influences other people to click on headlines*).

- five Likert scale responses (*Definitely yes, Probably yes, Might or might not, Probably not, Definitely not*)

Feedback: optional free text field for participants to leave feedback

Preliminaries of the survey included an information sheet and a consent form. We asked the participants for basic demographics information (age group, gender, country of residence, native language) and news reading habits (frequency of reading news).

The first part of the main survey consisted of 48 headlines (grouped into six groups of eight headlines for readability). The participants were asked: *How likely is it that other people will click on this headline?* Next to each headline was a single choice Likert scale. There were five Likert scale responses (*Extremely likely, Slightly likely, Neutral, Slightly unlikely, Extremely unlikely*).

The second part of the main survey consisted of 12 short sections for news values and style. The only omission from the features we presented in Chapters 4 and 5 is the news value of Uniqueness (one feature). It was not included because we decided to focus on news values which are expressed within a single headline, whereas the Uniqueness feature requires comparing headlines. In each of the twelve sections participants were presented with a short definition for the news value or style factor and one or more examples of its usage. Then they were asked the following questions: *"I personally consider this factor when clicking on headlines"* and *"I think this factor influences other people to click on headlines"*. For each they were given five Likert scale responses (*Definitely yes, Probably yes, Might or might not, Probably no, Definitely no*). We decided to include a question about subjective assessment of popularity for these factors, in order to investigate whether there are any differences between the two judgements.

In the final section participants had the option to comment on the survey and provide feedback using a free text field.

7.1.3 Participants and Quality Control

The crowdsourcing platform CrowdFlower was used to recruit participants for the survey. This allowed us to collect responses globally, thus reflecting the global nature of audiences of online news outlets. The survey took approximately 10 minutes to complete and participants were paid $2 for taking part. Out of 100 collected responses, 98 were recorded as complete in *The Guardian* version and 96 in *New York Times* version. While quality of responses was generally quite high, we carried out some quality control. We removed any responses where more than 75% of answers were neutral, as well as responses where time to complete was in the bottom quartile (to ensure that participants had taken time to understand the concepts). After the quality control measures, 71 responses were selected for *The Guardian* and 70 for *New York Times*. An overview of the survey participants, including their demographics, is presented in Table 7.1.

Table 7.1: Summary of survey participants

	The Guardian	*New York Times*
Recorded	100	100
Completed	98	96
After quality control	71	70
Age	<35: 47, >=35: 24	<35: 45, >=35: 25
Gender	Female: 17, Male: 54	Female: 13, Male: 57
Native English	Yes: 30, No: 41	Yes: 14, No: 56
News reading	Daily: 44, Weekly: 27	Daily: 52, Weekly: 18

For both *The Guardian* and *New York Times* surveys the majority of participants were under 35 (66% in *The Guardian*, 64% in *New York Times*), male (76% in *The Guardian*, 81% in *New York Times*), non-native speakers of English (58% in *The Guardian*, 80% in *New York Times*), and read news daily (62% in *The Guardian*, 74% in *New York Times*). This means that the survey was slightly biased towards male, non-native speakers of English. Since we do not have demographics information about the readership of *The Guardian* and *New York Times* we cannot say whether this sample differs from the users who have engaged with the headlines in our corpora.

7.1.4 Feedback from Participants

At the end of the survey we included a free-text field for feedback and comments. 28 participants of *The Guardian* version and 38 participants of the *New York Times* version filled out the field. We present examples of the feedback we received which provides some insight into our study objectives.

- surprise about the factors that might influence readers
  - "I never realized until I did this task that we(I) are so easily influenced(manipulated). Thank you"
  - "It hadn't occurred to me that grammar structure in headlines could be so important."

- comments about the extra-linguistic factors that impact news article popularity
  - "The size of the letters, the position on the page and the graphic material also have much influence"
  - "i think it's more thematics that attract people including me to news"
  - "[...] category related to their interests [...]"
- role of headlines
  - "I most of the time just go through only the headline because the headline itself gives you the full news. You don't have to read the full news"

The first examples show that some participants were not aware of the different factors which could influence headline popularity. The word *manipulated* was quite interesting, because of its negative connotation. Although we do not address it in this thesis, news outlet agenda would influence how headlines are written, e.g. optimising them for search engines[1] and thus 'manipulating' readers. Some participants also mentioned certain extra-linguistic factors (presentation and topic). Controlling the presentation factors was one of the objectives of this survey, and we conduct topic control in Section 8.2.3. We were happy to see one of the motivations for our thesis (importance of headlines) mentioned by a survey participant. Overall, the survey feedback provided us with several useful comments, but in order to gain more insight into how decisions were made when judging headline popularity, we conduct a qualitative analysis of survey responses by three experts in Section 7.3.

7.2 Headline Popularity for the Crowd

In order to better understand the impact of news values and style features on social media popularity (cf. Chapter 6), we used perceived popularity obtained from the survey as a popularity measure. The survey controls for factors like presentation and social influence, which means that the perceived popularity measure can provide clearer insights into popularity of headlines. Perceived popularity also targets slightly different aspects of popularity – clicking, rather than sharing on social media. Comparison of the two popularity measures can add new insights into the impact of news values and linguistic style on popularity. Since the collected responses are all categorical and we require a numeric measure to run the correlations, we decided to use the number of likely ratings (*Extremely likely* or *Slightly likely*) as the target popularity measure. These two responses reflect whether a reader would click on a headline mirroring the social media popularity

[1] This is very explicit in the *Yahoo!* style guide Barr (2010) where search engine optimisation is named as one of the objectives when formulating headlines

which reflects whether a reader has shared a news article on social media. The results are reported in Table 7.2.

The most noticeable result is that far fewer features significantly correlate with the perceived popularity measure compared to the correlations results for social media popularity in Chapter 6. In *The Guardian* only three features (polarity, difficult words, and verb phrase count) have statistically significant correlations. In *New York Times* that is the case for 13 features. We would like to note that in most cases there were not enough positive examples in the survey headlines to obtain statistics for binary features like modality or quote marks, hence the missing values.

Although fewer features correlate significantly with perceived popularity, for the ones that do the correlation is considerably stronger than in the case of social media popularity – the highest correlations we observed is τ=0.36 for parse tree height (compared to the strongest correlation with social media popularity which was τ=0.2). The statistically significant correlations also to a certain extent confirm our findings from Chapter 6. In particular, the strong correlations for Brevity and syntactic Simplicity feature are repeated for *New York Times*. The reason it is these features which correlate with both types of popularity might be that informativeness (i.e. longer, sentence-like headlines) is preferred by readers both when choosing to click on a headline, as well as when choosing to share it on social media. Features which significantly correlate with both social media popularity and perceived popularity will be particularly important for content creators who want to target both types of popularity.

7.3 Qualitative Analysis of Headline Popularity

In order to gain deeper insights about the choices that readers make about headlines, we conducted an in-depth qualitative analysis of surveys completed by expert annotators. The number of participants for the qualitative analysis was three: the author and two other researchers. Two of the three participants were non-British nationals residing for an extended period of time in the UK and regular readers of *The Guardian*. The third participant was a US national and regular reader of *New York Times*. Each participant independently completed the survey.

The responses for each headline were then discussed together, focusing in particular on agreements and disagreements. In the following sections we give examples which illustrate some interesting patterns and discuss them.

Table 7.2: Feature correlations with perceived popularity obtained from crowdsourced survey (N=48). For numeric features we used Kendall's τ correlation. For binary features we checked whether the feature median is statistically different from the overall median (*The Guardian*: 33, *New York Times*: 34.5) using Wilcoxon signed rank test.

	The Guardian		*New York Times*	
Feature	Statistic	Sig.	Statistic	Sig.
News values				
number of entities	0.16	0.14	-0.06	0.61
Wikipedia burst size	-0.08	0.49	0.09	0.46
Wikipedia burstiness	0.08	0.46	0.10	0.34
Wikipedia long-term prominence	0.15	0.16	0.17	0.11
Wikipedia day-before prominence	0.09	0.40	**0.23**	**0.03**
news recent prominence	0.11	0.34	**0.27**	**0.02**
sentiment	0.13	0.20	-0.03	0.82
polarity	**0.24**	**0.02**	0.20	0.07
connotations	-0.03	0.75	0.08	0.47
bias	0.14	0.19	**0.32**	**0.00**
comparative/superlative	0.01	0.92	0.18	0.15
intensifiers	-0.06	0.63	0.04	0.72
downtoners	NA	NA	0.14	0.25
proximity	31.00	1.00	36.00	0.69
uniqueness	0.05	0.70	0.16	0.17
surprise	0.02	0.82	-0.03	0.83
Linguistic style				
number of words	0.06	0.60	**0.27**	**0.01**
number of characters	0.02	0.83	**0.27**	**0.01**
parse tree height	0.09	0.40	**0.36**	**0.00**
non-terminal nodes	0.08	0.45	**0.30**	**0.00**
entropy	-0.11	0.25	-0.10	0.31
difficult words	**-0.22**	**0.03**	-0.15	0.16
information content	0.17	0.08	0.04	0.72
word frequency	0.17	0.10	**0.29**	**0.00**
number of senses	0.13	0.20	-0.02	0.89
modality	35.00	1.00	NA	NA
exclamation mark	NA	NA	30.00	1.00
question mark	39.00	0.89	21.00	1.00
quote marks	41.50	0.80	NA	NA
three consecutive nouns	37.00	0.79	35.00	1.00
NP count	-0.15	0.15	**-0.32**	**0.00**
proportion of nouns	0.02	0.82	0.20	0.07
proportion of proper nouns	0.10	0.36	**0.25**	**0.02**
VP count	**0.20**	**0.05**	**0.29**	**0.01**
proportion of verbs	0.16	0.14	**0.22**	**0.05**
proportion of adverbs	-0.11	0.35	**0.27**	**0.02**

7.3.1 Analysis of Disagreements

We looked at cases where the annotators chose different responses, in particular where the differing responses where at the opposite ends of the scale (i.e. positive vs. negative). We identified several sources of disagreements in our responses which are described below.

Familiarity with the news source. When asked to judge whether other people would click on a headline, one annotator interpreted the instructions to mean "an average person like her", whereas another annotator was thinking of the majority of readers for a given news source. For example: one annotator familiar with *The Guardian*'s readership judged the Headlines 7.1 and 7.2 to be relevant for most readers, but another annotator less familiar with that news source judged these headlines as not relevant for an average reader. As we are looking at news outlets from two countries, familiarity with the news source also had a bearing on familiarity with the entities present in the headlines, which would impact the news values of Prominence and Proximity.

Headline 7.1. "Chicken schnitzel with herbs and parmesan – Bondi Harvest video recipe" (*The Guardian*)

Headline 7.2. "Ask a grown-up: how does a squid make ink?" (*The Guardian*)

Familiarity with the domain. Familiarity with various topics differed among the annotators and without understanding the headline they could not make an accurate judgement about whether other people would click on a headline. Headline 7.3 is an example of how background knowledge (or lack of it) influenced the annotators' choices. In this particular case background knowledge is related to Proximity – the annotators that reside in the UK had at least a passing familiarity with the entities which occurred in the headline.

Headline 7.3. "Wigan v Arsenal: FA Cup semi-final – as it happened" (*The Guardian*)

Vague headlines and lack of context. Another issue that we identified was the lack of sufficient detail in the headline to understand the content. As a result of that, the annotators had to make an educated guess. For example, Headlines 7.4 and 7.5 include abstract phrases and no specific entities. For some annotators that lack of specificity was judged negatively, but for another it did not influence their decisions. We note that outside of a survey setting where only the headline text was presented, more context might be available (e.g. browsing a particular section, or an image or a thumbnail accompanying the headline), which can help to disambiguate the meaning of the headline.

Headline 7.4. "The Meaning of Fulfillment" (*New York Times*)

Headline 7.5. “A Rational Quarantine” (*New York Times*)

Overall, most of the disagreement between the annotators arose because of difficulties with understanding the headline. These difficulties were due to lack of familiarity with the news source and with the entities that appeared there, lack of familiarity with the domain, and lack of context. We address the first issue in Chapter 9 where we build country-specific models of social media popularity and add country relatedness to the implementation of Prominence and Proximity. We address the second issue in Chapter 8 where we build prediction models for two topics separately, which provides topic control. We do not address the third issue directly, however our findings on the impact of style features strongly suggest that longer headlines which provide more information are more likely to be popular.

7.3.2 Analysis of Positive Agreements

Next we look at the cases where the three annotators chose a positive response. These cases point to factors which strongly influence popularity and help us to interpret the impact of our features.

Use of news values. The annotators agreed about the perceived popularity of Headlines 7.6 and 7.7, because they mentioned prominent entities (HARVARD, SECRET SERVICE) and were quite surprising. This points to the importance of the news values of Prominence and Surprise. Headline 7.8 was thought to be quite eye-catching, because of the ‘top 10’ phrase, as well as the surprising phrase ‘daftest ways to become a world champion’. Both ‘top’ and ‘daftest’ fall under the news value of Magnitude and our implementation captures them. On the other hand, Headline 7.9 also features the news value of Magnitude (‘47 Years’). In this case our implementation would not identify it since the magnitude is relative (i.e. 47 years is considered a long span in this context, but perhaps not in others). This type of relative Magnitude would need to use world knowledge in the implementation.

Headline 7.6. “From a Rwandan Dump to the Halls of Harvard” (*New York Times*)

Headline 7.7. “The Collapse of the Secret Service” (*New York Times*)

Headline 7.8. “Top 10 daftest ways to become a world champion” (*The Guardian*)

Headline 7.9. “An Apple a Day, for 47 Years” (*New York Times*)

Use of linguistic style. The linguistic style in Headline 7.10 was what caught the attention of the annotators. The linguistic style features for this headline which were found to correlate positively with social media popularity (cf. Table 6.1) and were above the feature median (cf. Table 5.2) were: number of words and syntactic simplicity features (parse tree height and number of non-terminal tree nodes). Although journalistic literature and style guides recommend short, simple headlines, we found a statistically significant positive correlation for these features. The annotators' positive opinion about this headline (that it is informative, the use of fronting) helps to explain our findings. We also found that for *The Guardian* all genres punctuation was associated with higher social media popularity. The annotators agreed that the use of punctuation in Headlines 7.11 and 7.12 is what drew their attention. In Headline 7.11 the quote marks prompted questions such as "How exactly children were kept from parents and why?", whereas for Headline 7.12 the use of the question was to directly engage with the reader ("when did *you* realise"). Both headlines offer a possible explanation for the positive result for Punctuation features in *The Guardian*.

Headline 7.10. "No-fly list used by FBI to coerce Muslims into informing, lawsuit claims" (*The Guardian*)

Headline 7.11. "Children 'kept from parents' at centre for failed asylum seekers" (*The Guardian*)

Headline 7.12. "Equal pay awakenings: when did you realise you were underpaid?" (*The Guardian*)

Overall, we noted that when the annotators agreed on a positive response, the headline included at least one feature which we found to have a positive correlation with popularity in Chapter 6. These cases show that the annotators agreed despite their individual differences and backgrounds (which we noted when discussing disagreements). This makes a case for the features we implemented and helps to explain the unexpected behaviour of some features.

7.3.3 Influencing Factors

There were three factors which we observed to influence the annotators' choice on whether a headline would be clicked on or not.

Firstly, we noted audience **familiarity** effect in the choices that the annotators made. All annotators said they were accustomed to particular newspaper styles and reading headlines in a very different style made parsing the content more difficult. This was particularly in evidence when the two UK residents made the same choice and the US resident disagreed.

Secondly, in cases where the **genre** of the article is clear from the headline, the genre did affect how the headlines was interpreted by the annotators. For example, because

Headlines 7.13 and 7.14 are headlines of review articles the sentiment in the headlines ('badness', 'hard-working') did not seem to have as much an effect. This was also noted by the annotators of news values in headlines (cf. Section 4.2).

Headline 7.13. "Stinkbomb and Ketchup-Face and the Badness of Badgers by John Dougherty" (*The Guardian*)

Headline 7.14. 'Bodies of Light by Sarah Moss review – 'a hard-working novel about hard-working women"' (*The Guardian*)

Finally, it was clear that there were headlines which would be interesting to very **specific audiences**. For example, Headlines 7.15 and 7.16 would be of interest to fans of the relevant sport, but probably not to wider audiences. Similarly, Headline 7.17 would most likely be clicked on by readers with an interest in politics and/or economics, but especially from an American perspective this headline might be quite esoteric. These examples make a good case for topic control and it is one of the prediction experiments we conduct in Chapter 8. It also indicates that taking into account user preferences (via a user model) might improve the model for news article popularity.

Headline 7.15. "Two Danish Badminton Players Report a Fixing Invitation" (*New York Times*)

Headline 7.16. "N.B.A. to Experiment With Shorter Game in Nets Exhibition" (*New York Times*)

Headline 7.17. "Lawmakers Grill French Candidate for European Economic Post" (*New York Times*)

7.4 Judgements about News Values and Style

Results of Part 2 of the survey (Judgement criteria) are presented in Figures 7.1 (personal judgement) and 7.2 (judgement about other people). Responses from the five-point Likert scale were classified as positive (*Definitely yes* and *Probably yes*), neutral (*Might or might not*), and negative (*Probably not* and *Definitely not*).

For personal judgements (Figure 7.1) the percentage of positive ratings ranged from 78% (Proximity) to 32% (Nouns). The highest percentages of neutral and negative ratings were for Nouns and Verbs (40% and 28% negative in both cases). For judgements about other people's preferences (Figure 7.2) the percentage of positive ratings ranged from 92% (Prominence) to 42% (Verbs). The highest percentage of neutral ratings was also for Verbs (40%). The highest percentage of negative ratings was for Nouns (28%). The highest

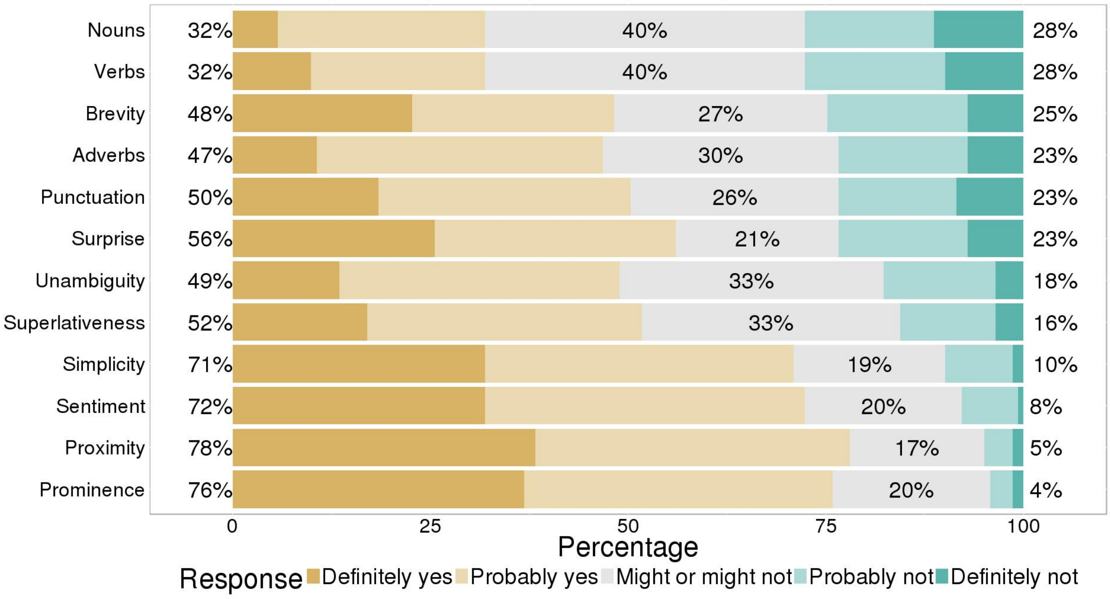


Figure 7.1: Judgements (*"I personally consider this factor when clicking on headlines"*) about news values and style (combined *The Guardian* and *New York Times*, N=141). Percentages refer to combined positive (*Definitely yes* and *Probably yes*), neutral (*Might or might not*), and negative (*Probably not* and *Definitely not*) judgements.

rated (i.e. with the highest percentage of positive responses) news values were Proximity (personal: 78%, other people: 83%), Prominence (personal: 76%, other people: 92%), and Sentiment (personal: 72%, other people: 84%). The highest rated style factors were Simplicity (personal: 71%, other people: 72%) and Punctuation (personal: 50%, other people: 56%).

News values vs. style. For both personal judgements and judgements about other people there is a clear pattern of news values achieving high ratings and style factors achieving more mixed ratings. For news values the percentage of positive ratings ranged from 56% (Surprise) to 78% (Proximity) for personal judgements; and from 72% (Surprise) to 92% (Prominence) for judgements on other people's preferences. For style the percentages of positive ratings are much lower: between 32% (Nouns and Verbs) and 50% (Punctuation) for personal judgements; and between 42% (Verbs) and 72% (Simplicity) for judgements about other people's preferences. Style factors also received higher percentages of neutral responses (up to 40% in both cases) than news values (up to 33% for personal and up to 23% for other people). In both cases only news values achieved more than 50% positive responses. This difference in the responses between news values and style might be because higher level concepts like Prominence or Proximity are more salient to readers than relatively technical concepts like Nouns or Adverbs. While style might not be

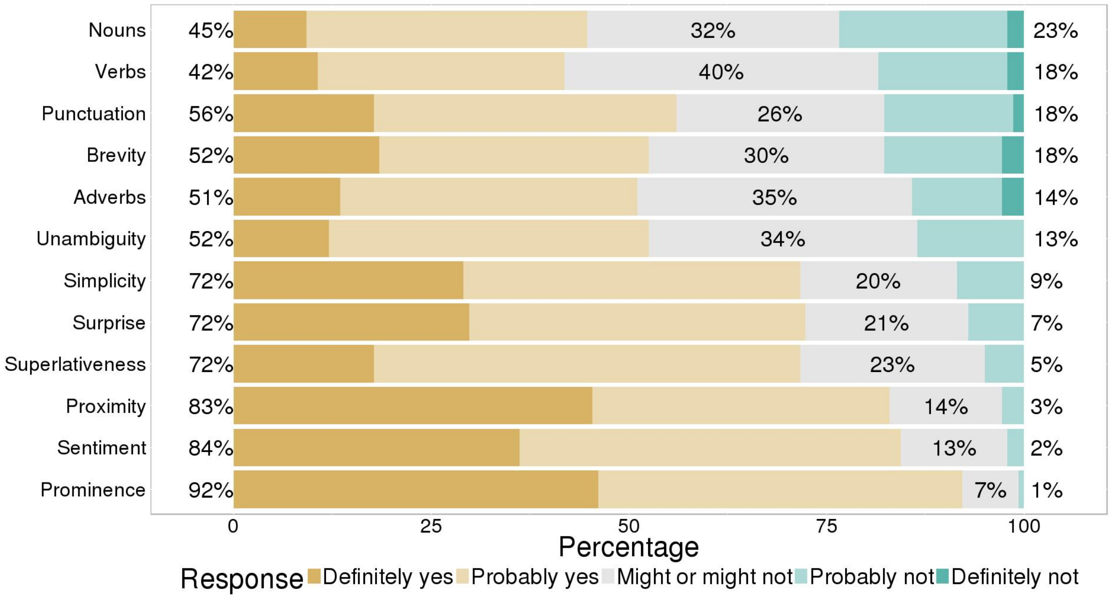


Figure 7.2: Judgements (*I think this factor influences other people to click on headlines*) about news values and style (combined *The Guardian* and *New York Times*, N=141). Percentages refer to combined positive (*Definitely yes* and *Probably yes*), neutral (*Might or might not*), and negative (*Probably not* and *Definitely not*) judgements.

perceived by readers to influence their choice of headlines to a great extent (since headlines in news outlets should already be grammatical), we found almost all linguistic style features to correlate significantly with social media popularity (cf. Chapter 6). One of the comments from survey participants (cf. Section 7.1.4) also indicated that some readers might not realise that they are influenced by the linguistic style of headlines.

Personal vs. other people. In general personal judgements and judgements about other people's preferences tended to have similar percentages of negative responses. However, personal judgements had lower percentages of positive responses and higher percentages of neutral responses (especially for news values). For example, the news value of Surprise had 72% positive responses when judging for other people, compared to 56% when judging personally. This follows some studies which show how bias is perceived to affect oneself less than to affect other people (Pronin et al., 2004).

Survey judgements vs. impact on social media popularity. We observed some interesting differences between what headline aspects people say they consider when choosing to click on headlines compared to our experimental results in Chapter 6. Overall, survey participants judged news values more positively than style features, however in the correlation results for social media popularity (cf. Tables 6.1 and 6.2) style features achieved some of the highest correlations of all features. Indeed, some of the style features that had high correlations with social media popularity (e.g. Brevity, Nouns) had high percentages of neutral or negative responses in the survey. These disparities might be due to a difference in the engagement with the headline – the survey asked about clicking on a headline (personal, direct engagement), whereas the correlations target explicit social behaviour such as likes or retweets. Moreover, news values are high-level concepts which are quite easy to understand given some examples. On the other hand, some style features (Nouns, Verbs, Adverbs) require some knowledge about parts of speech which an average reader may not have, leading to them not realising these are factors that actually influence them.

7.5 Summary

In order to model the social media popularity of news articles using headlines we need to know what impact individual headline features have on popularity. In the previous chapter we found that most of the news values and linguistic style features we propose for this task correlate significantly with popularity of news articles on Twitter and Facebook. To gain deeper insights about these findings and to control for confounding factors like headline presentation and social influence, in this chapter we presented the results of a crowdsourced survey which gave us a measure of perceived popularity of headlines. We used that measure to correlate with feature values and compare the results with our findings on correlations with social media popularity. The qualitative analysis of the survey responses also helped to explain some of our findings in the previous chapter.

We found that although there were fewer statistically significant correlations with perceived popularity compared to social media popularity, the correlations were considerably stronger. The features that had significant correlations with both perceived and social media popularity measures (e.g. Brevity, syntactic Simplicity) are the ones that are crucial for content authors to implement.

The qualitative analysis of survey responses by three experts yielded several insights. Familiarity with the news source style and individual topic preferences influence readers. We aim to control for these factors in two ways. Firstly, we conduct within-topic prediction experiments (cf. Section 8.2.3) in order to minimise the effects of topic preferences. Secondly, we address the news source familiarity issue by building country-specific prediction models in Chapter 9. Readers located in UK will be more familiar with *The Guardian* (at least in passing) than readers outside UK, which will allow us to account for differences in headline style.

The second part of the crowdsourced survey asked for judgements about the effect of news values and style features when choosing to click on a headline. We observed some interesting discrepancies from our findings in Chapter 6. Survey participants judged news values to have very positive impact, however the correlations we obtained for social media popularity were higher for linguistic style features.

Our main contributions for this chapter are:

(i) obtaining perceived popularity measure using a crowdsourced survey

(ii) comparing news values and linguistic style correlations with social media popularity and perceived popularity

(iii) qualitative analysis of expert annotators' survey responses

(iv) analysis of what readers judge to influence their decisions when clicking on headlines

In this and the previous chapters we have shown that news values and style have a significant impact on social media popularity, and are judged to affect choice of headlines by individual readers. These findings indicate how to reformulate a headline to achieve higher social media popularity. However, in order to know when such a reformulation is advisable, we first need to be able to get an expectation of the response a headline will get on social media. To do that in the next chapter we build prediction models which use news values and linguistic style features we proposed.

Chapter 8

Social Media Popularity Prediction Using News Values and Style

In this thesis we model the social media popularity of news articles using headline-derived news values and linguistic style features. In previous chapters we showed that journalism-inspired news values can be operationalised for automatic extraction from headlines (Chapter 4). We also showed the implementation of features related to the linguistic style of headlines (Chapter 5). In order to model the social media popularity of news articles we needed to find out how these features correlate with popularity *individually*. For both news values and linguistic style features we found significant correlations with social media popularity (Chapter 6) and to a lesser extent with perceived popularity (Chapter 7). The qualitative analysis of perceived popularity helped to explain some of our findings on social media popularity. These earlier chapters provide insights about how individual news values and linguistic style features correlate with social media popularity. These findings can be used to inform the *formulation* of headlines. However, in order to know when a headline needs to reformulated, the headline author first needs to have an *expectation of social media popularity for a news article*. To address that need in this chapter we build prediction models of social media popularity of news articles using news values and linguistic style features of headlines.

In this chapter we describe the method and present the results of using news values and linguistic style features in a social media popularity prediction model for individual news articles from *The Guardian* and *New York Times*. The goal of this model is to predict the social media popularity of any news article from a given news source from the article's headline. This is in contrast to approaches which try to predict whether something will become viral on social media, e.g. Berger and Milkman (2012), or approaches which primarily aim to rank news articles (Tatar et al., 2014). Our motivation for this choice is that authors of many news articles will not be expecting their content to become viral or be highly ranked. Instead they will be looking for an indication of popularity (the prediction models we build in this chapter) and a suggestion on how to improve it (the correlations with popularity in Chapters 6 and7). Ours is also the first approach to use solely the headline text to make a prediction of social media popularity for news articles.

8.1 Method

We use regression to predict the popularity of news articles on social media. While regression is a more difficult task compared to classification, it also provides a more fine-grained prediction, which is more suitable for the creative writing support setting we envisage as a possible application of this research (cf. Section 1.1). Moreover, Arapakis et al. (2014) argued against using classification for popularity prediction, due to the arbitrary class splits potentially introducing bias towards articles with low popularity.

Source control. We adopt source control in our approach, which helps us to investigate the predictive power of the features we proposed. *The Guardian* and *New York Times* are both major global news outlets, however there is a considerable difference in their readership size (cf. Section 3.1), which might influence to what extent their content is accessed and shared. When a news source has been used as a feature (Bandari et al., 2012; Arapakis et al., 2014) it has been found to be a strong predictor of news article popularity. However, from a news outlet perspective such a feature is not useful, since they cannot change that aspect of their news production. On the other hand, characteristics of headline text like news values and style can be edited based upon findings on the impact of individual features on popularity. That is why our prediction models were built for each source separately, thus avoiding the effect of source popularity.

Algorithm choice. We experimented with a number of regression models (linear regression, decision tree, support vector machines) and found that the best results were achieved using support vector regression with RBF kernel (Chang and Lin, 2011). The results reported in this chapter all use support vector regression. The popularity measures – T and F (Twitter and Facebook popularity respectively) – were log-transformed in order to improve model fit.

Evaluation. The prediction models were evaluated on a separate test set for each news source, as opposed to using cross-validation. Two factors determined this decision: (i) evaluation using cross-validation is not appropriate, because the data is temporally ordered and the training data should always temporally precede the test data; and (ii) one of our features, Uniqueness, makes use of the temporal ordering of headlines. Two evaluation metrics were used: Kendall's rank correlation coefficient (τ; Kendall (1938)) and mean absolute error (MAE). Significance testing was performed using z-test for the correlations and t-test for the errors.

Baselines. We implemented three baselines: one standard unigrams baseline and two reimplementations of state-of-the-art 'cold-start' approaches to news article popularity prediction. We present a brief overview of the baselines notation in Table 8.1 and description of the baselines below. Our model is denoted as $\mathcal{M}$.

Table 8.1: Overview of baselines.

Baseline notation	Baseline description
$\mathcal{M}$	Our full model
$\mathcal{M}_{\mathcal{U}}$	Unigrams baseline
$\mathcal{M}_{\mathcal{B}}$	Reimplementation of Bandari et al. (2012)
$\mathcal{M}_{\mathcal{A}}$	Reimplementation of Arapakis et al. (2014)

The first baseline is unigrams ($\mathcal{M}_{\mathcal{U}}$). We used 1000 most frequent unigrams, excluding stopwords.

In Section 2.2.2 we presented on overview of approaches to news article popularity prediction. We identified two approaches to compare against: Bandari et al. (2012) ($\mathcal{M}_{\mathcal{B}}$) and Arapakis et al. (2014) ($\mathcal{M}_{\mathcal{A}}$). A feature-by-feature comparison of the models is provided in Appendix F. Although originally both approaches used full article text, we ran these baselines on the same dataset as ours (that is to say, headlines only). We aimed at as close reimplementation as possible, but in some cases we had to make adjustments. We used Stanford Named Entity Recognizer (Finkel et al., 2005) and SentiWordNet (Baccianella et al., 2010) for Prominence and Sentiment features respectively. As we do not have access to archival Twitter data, we used Wikipedia to calculate Prominence features. Finally, unlike in the original implementations there is no news source feature, because we conduct a source-internal evaluation.

The two previous approaches we reimplement, as well as similar tasks (Lakkaraju et al., 2013), make use of some of the metadata that is available at the time of article publication (such as article category, or publication time). The reimplemented baselines ($\mathcal{M}_{\mathcal{B}}$ and $\mathcal{M}_{\mathcal{A}}$)

and our full model ($\mathcal{M}$) also include metadata. Following the implementation by Arapakis et al. (2014), both article category and publication date and time are implemented as binary features in our model. In case of *New York Times* articles, the category is indicated by both genre and section information and we make use of both as binary features. The Bandari et al. (2012) baseline calculates a category score.

8.2 Prediction Results

In this section we present prediction results for three experimental settings. Firstly, we start with comparing the performance of our model against baselines to see if our approach offers an improvement over the state of the art. Secondly, we investigate what is the predictive power of headline-derived news values and linguistic style by using different feature subsets for prediction. Finally, we build prediction models which are controlled for news article category. This is in order to drill down into the effect of news values and linguistic style features without the confounding factor of category.

8.2.1 Performance against Baselines

The performance of our prediction model is compared against baselines in two settings: (i) using all features in Table 8.2, and (ii) using headline-derived features only in Table 8.3. This allows the exploration of the predictive power of headline-derived features for prediction of news article popularity, and to examine the benefit of a model consisting of features which we introduced.

Table 8.2: Prediction results of our model ($\mathcal{M}$) against baselines using all features (news values, style, metadata). Result in bold indicates improvement significant at p<0.05.

	The Guardian				*New York Times*			
	τ		MAE		τ		MAE	
	T	F	T	F	T	F	T	F
$\mathcal{M}_\mathcal{U}$	0.32	0.25	0.82	1.59	0.19	0.22	**0.66**	1.68
$\mathcal{M}_\mathcal{B}$	0.36	0.29	0.71	1.53	0.15	0.18	0.67	1.72
$\mathcal{M}_\mathcal{A}$	0.41	0.35	0.7	1.45	0.21	0.3	0.86	1.57
$\mathcal{M}$	**0.43**	**0.37**	**0.68**	**1.42**	**0.23**	**0.32**	0.88	**1.54**

When using all features (news values, style, metadata; Table 8.2), our model ($\mathcal{M}$) significantly outperforms all baselines both in terms of the correlation results and the errors. The only exception is the MAE result for the unigrams baseline for Twitter popularity in

New York Times dataset. The highest correlation ($\tau = 0.43$) was achieved for the Twitter popularity measure in *The Guardian* dataset, while the lowest error was achieved for the one unigrams baseline result which outperformed our model. In general our model achieved greatest gains over the unigrams baseline, and the smallest gains over the baseline which reimplemented the approach by Arapakis et al. (2014). Both the correlations and errors results were better for *The Guardian* than *New York Times*.

Table 8.3: Prediction results of our model ($\mathcal{M}$) against baselines using headline features only (news values and style). Result in bold indicates improvement significant at *p*<0.05.

	The Guardian				*New York Times*			
	τ		MAE		τ		MAE	
	T	F	T	F	T	F	T	F
$\mathcal{M}_{\mathcal{B}}$	0.11	0.07	0.94	1.74	0.05	0.02	0.7	1.85
$\mathcal{M}_{\mathcal{A}}$	0.22	0.19	0.88	1.66	0.19	0.16	**0.67**	1.75
$\mathcal{M}$	**0.29**	**0.26**	**0.83**	**1.59**	**0.21**	**0.23**	0.69	**1.66**

When using only features which were derived from headline text (i.e. dropping the metadata features; cf. Table 8.3), the prediction performance drops considerably. Now the highest correlation is still for Twitter popularity in *The Guardian* dataset, but without metadata features it reaches $\tau = 0.29$. The lowest error (MAE=0.67) is also achieved for the Twitter popularity measure, but in *New York Times* dataset. It is the only measure where a baseline (in this case, $\mathcal{M}_{\mathcal{A}}$) outperforms our model. As with the prediction results which used all features, the performance is better in *The Guardian* than *New York Times* dataset.

8.2.2 Effect of Feature Groups on Prediction Performance

In this section we examine the performance of four feature groups: news values ($\mathcal{M}_{\mathcal{N}}$), linguistic style ($\mathcal{M}_{\mathcal{S}}$), all headline features (i.e. news values and linguistic style combined, $\mathcal{M}_{\mathcal{N}+\mathcal{S}}$), and metadata ($\mathcal{M}_{\mathcal{M}}$). This is in order to establish their prediction performance, which can inform the implementation of certain feature groups for the task of news article popularity prediction and for augmenting other approaches with our features. We report the results of the feature group comparison in Table 8.4.

The full model which uses all available features achieves the best performance overall. The feature group with the highest overall performance is metadata. This is not unexpected, since we observed some article categories or genres (which are included in the metadata features) attracting very high levels of social media attention (cf. Tables 3.6 and 3.7 in Chapter 3). The feature group with the lowest performance is news values. Although we

Table 8.4: Prediction results by feature group: $\mathcal{M}_{\mathcal{N}}$ = news values, $\mathcal{M}_{\mathcal{S}}$ = style, $\mathcal{M}_{\mathcal{N}+\mathcal{S}}$ = news values and style, $\mathcal{M}_{\mathcal{M}}$ = metadata, $\mathcal{M}$ = all features. Result in bold indicates improvement significant at *p*<0.01.

	The Guardian				*New York Times*			
	τ		MAE		τ		MAE	
	T	F	T	F	T	F	T	F
$\mathcal{M}_{\mathcal{N}}$	0.2	0.17	0.89	1.67	0.14	0.14	**0.68**	1.74
$\mathcal{M}_{\mathcal{S}}$	0.25	0.22	0.86	1.62	0.18	0.19	0.7	1.7
$\mathcal{M}_{\mathcal{N}+\mathcal{S}}$	0.29	0.26	0.83	1.59	0.21	0.23	0.69	1.66
$\mathcal{M}_{\mathcal{M}}$	0.39	0.33	0.72	1.51	0.17	0.23	0.92	1.65
$\mathcal{M}$	**0.43**	**0.37**	**0.68**	**1.42**	**0.23**	**0.32**	0.88	**1.54**

observed some of the higher correlations with social media popularity for a number of individual features (cf. Sections 6.1 and 6.2), the group as a whole performs worse than the others. On their own, news values features reached a correlation of up to $\tau = 0.2$. It was particularly notable that style features achieved a correlation as high as $\tau = 0.25$ on their own, considering that they do not convey any topical information at all and focus solely on wording. This might be because linguistic style features occur in most headlines, so the features which correlate highly and have high prevalence (e.g. Brevity, syntactic Simplicity) aid the prediction performance. We also use all features derived from the headline text (i.e. news values and linguistic style) and find that it outperforms the news values and linguistic style models. Crucially, it matches and even outperforms the metadata features in *New York Times* corpus. This is a strong indicator that headline content (beyond simply the category) can be used as a predictor of social media popularity of news articles.

8.2.3 Effect of Category on Prediction Performance

In the prediction models reported so far we already controlled for the news source. That is to say, we built the prediction model for both news sources separately and did not use the news outlet as a predictor. This allowed us to focus more on the effect of the headline features, without the confounding factor of the popularity of the news source. Similarly, in this section we added another step, in order to remove another confounding factor – that of the news article category. In Table 8.4 we observed that metadata features achieve the overall best performance of all feature groups. This can be linked to the predictive power of topic and genre information. This aligns with earlier research about the varying levels of social media popularity for different topics and genres (Bastos, 2015).

In order to control for the popularity of category information, we ran the prediction models using only a subset of the corpus from a particular category. To continue with source control, we chose two categories (*World news* and *Sports*), because they appear in both *The Guardian* and *New York Times* datasets and have enough articles to train the prediction models. As we are using headlines from only one category, we make use of only headline-derived features (model $\mathcal{M}_{\mathcal{N}+\mathcal{S}}$ in Table 8.4). Results are reported in Table 8.5 (for *The Guardian*) and Table 8.6 (for *New York Times*).

Table 8.5: Category-control prediction results for *The Guardian*. We also report the number of articles in training and test sets for each category.

			τ		MAE	
	No. training	No. test	T	F	T	F
World news	1854	2233	0.06	0.03	0.62	1.59
Sports	901	1247	0.11	0.15	0.63	1.2
All	11980	13806	0.29	0.26	0.83	1.59

Table 8.6: Category-control prediction results for *New York Times*. We also report the number of articles in training and test sets for each category.

			τ		MAE	
	No. training	No. test	T	F	T	F
World news	662	711	0.1	0.18	0.54	1.44
Sports	634	732	0.24	0.3	0.47	1.29
All	5074	5011	0.21	0.23	0.69	1.66

The category-control predictions yielded some interesting results. In both datasets the correlations for individual topics (*World news*, *Sports*) were lower than the full model (*All*), with the exception of the correlation results for *Sports* in *New York Times* where the category-controlled model achieved higher correlations. In case of *The Guardian* the drop in the correlation results was very noticeable, especially for topic *World news* – from τ=0.29 and τ=0.26 to τ=0.06 and τ=0.03 for Twitter and Facebook measures respectively. While there was a drop in performance for correlations, the errors were actually lower for the category-controlled models. This might be because the size of the error is sensitive to the relative popularity of a given category. We noted in Chapter 3 (Tables 3.6 and 3.7) that categories differ quite widely in popularity and we repeat that analysis for categories *World news* and *Sports* in Table 8.7. We believe that the lower standard deviation in popularity might lead to smaller errors.

Table 8.7: Comparison of topic popularity for selected topics in *The Guardian* and *New York Times*. Median value and standard deviation (SD) of popularity metrics is reported.

	The Guardian				*New York Times*			
	The Guardian				*New York Times*			
	$T(SD)$		F	$F(SD)$	T	$T(SD)$	F	$F(SD)$
World news	61	87.03	79	3376.5	94	157.89	264	10982.33
Sport	18	30.95	9	1072.09	158.5	94.16	35	2020.99
All	41	83.54	42	2694.25	102	280.16	153	10299.14

The results of the category-control prediction indicate that when trying to predict the exact amount of attention of social media, category control results in smaller errors. However, when the priority is ranking of news articles by popularity, then category-control prediction does not improve – and usually lowers – the performance of the prediction model.

8.3 Discussion

In this chapter we reported the regression results against baselines in Tables 8.2 and 8.3. We also looked at the performance of feature groups in Table 8.4 and at category control in Tables 8.5 and 8.6.

Performance against baselines. Our model significantly outperforms all baselines with the exception of MAE results for *New York Times*. Despite considering a number of factors (e.g. mistake in the code or in the input files), we could find no explanation for these. On the whole, this is a promising result, considering that headline text provides only a limited amount of data. While the $\mathcal{M}_{\mathcal{A}}$ baseline might achieve only slightly lower performance to our model with fewer features, it does not offer as many insights into how headline features impact an article's social media popularity. Excluding metadata features (which are the same for both $\mathcal{M}_{\mathcal{A}}$ and $\mathcal{M}$), our model ($\mathcal{M}$) has 36 features compared to 9 in the $\mathcal{M}_{\mathcal{A}}$ baseline, thus offering more dimensions for potential editing and improvement of the headline.

Performance of feature groups. Using all features significantly outperforms any individual feature group at p<0.01. Metadata (especially category) is the strongest feature group overall, suggesting that topic and genre of the article play a significant role for readers. However, the model which combines news values and linguistic style features matches or outperforms the metadata model, which indicates that headline content is useful for prediction. Furthermore, news values and style features lend themselves to editing by the headline author, while topic and genre are probably impossible to change. We note that although not all news values features are suitable for editing – the journalist cannot change the fact that a news event is about a certain entity, they can choose to refer to that entity explicitly (e.g. *Theresa May* instead of *Prime Minister*), thus raising the value for Prominence. While news values achieve the lowest performance of all groups, they include some of the features which we found to be best correlated (e.g. Prominence, Sentiment; cf. Tables 6.1 and 6.2 in Chapter 6), which makes them good candidates when reformulating a headline to achieve higher social media popularity. It is especially noteworthy that style features, which are largely topic-independent, on their own achieve good performance. This suggests that headline style is important to social media readers, independent of article content, which makes our proposed application of creative writing support even more beneficial.

Category-control performance. In the category control setting (i.e. predicting popularity for articles within news article category) we saw an improvement in MAE measure, as well as improvement in correlations results for category *Sports* in *New York Times*. The decrease in errors when controlling for category indicates that category control prediction could be useful when trying to minimise prediction errors, rather than accurately rank news articles.

8.4 Summary

In this chapter we presented the results of predicting the popularity of news articles on social media using news values and linguistic style features derived from headline text. We are the first to use headlines to predict the popularity of news articles on social media. Using our features we significantly improved over a unigrams baseline and reimplementations of state-of-the-art approaches by Bandari et al. (2012) and Arapakis et al. (2014).

This indicates that content of headlines does have a link with social media popularity, as we showed that headline text provides enough data to beat several baselines, which can inform guidelines about headline writing or real-time writing and editing support systems. Combined with our findings about news values and linguistic style features' significant correlations with social media popularity, we are able to make recommendations about reformulating headlines.

By conducting a source-internal evaluation (i.e. building a prediction model for each source separately) we gained clearer insights about the predictive power of feature groups. We found that although news article category is often the best predictor of popularity, the headline-derived features we propose also achieve significant results (even matching and outperforming metadata for *New York Times* corpus) leading to a correlation of 0.29 for headline-derived features only. We also observed differences in prediction results between the two news sources, whereby the prediction results are overall better for *The Guardian* compared to *New York Times*, which might be due to the fine granularity of categories.

Our main contributions for this chapter are:

(i) we built prediction models of news articles' social media popularity using headlines

(ii) we showed that news values and linguistic style features outperform state-of-the-art 'cold-start' baselines

(iii) we showed that in certain cases features derived from headline text match or outperform metadata features

(iv) we found that in certain cases category control can help to decrease prediction errors

In this chapter we presented prediction results for a global audience. However, in Chapter 6 we noted the unexpected lower popularity for headlines with the Proximity news value, and in Chapter 7 we heard from expert annotators that familiarity with the news source had bearing on their choice of headlines. These findings indicated that certain user characteristics (in particular user location) might influence the performance of our model. In Chapter 9 we investigate whether augmenting the Prominence and Proximity features to take into account the readers' location results in an improvement in prediction performance.

Chapter 9

Country-Specific Prediction Model

In this thesis we model the social media popularity of news articles using headlines. We proposed the use of news values (Chapter 4) and linguistic style (Chapter 5) features. We established that these features correlate significantly with social media popularity (Chapter 6), and compared these results with perceived popularity obtained through a crowdsourced survey (Chapter 7). In the previous chapter we combined these features to build prediction models for global social media popularity. In these earlier investigations we noted that the location of the user can influence the results of the model. We hypothesised that the lower popularity of headlines with the Proximity news value (cf. Chapter 6) could be due to our not controlling for user location (cf. our first implementation of Proximity in Section 4.1.4, where we took the majority of the readership for a given news source to be from the same country where the news outlet is published). This was confirmed through a qualitative analysis of experts' survey responses on headlines' perceived popularity, where we observed that familiarity with the news source through being resident in the country of the news source allowed the annotators to recognise certain country-related entities and thereby judge their Prominence. As the implementations for the Proximity and Prominence are *location-agnostic* (i.e. we are not controlling the reader location) we cannot draw firm conclusions about geographic relevance and its impact on social media popularity.

In this chapter we will address this issue by using *country-specific*, instead of global popularity. We do this by geolocating Twitter users who have tweeted or retweeted one of the news articles in our corpora. Once we have the country-specific popularity we (i) augment Proximity and Prominence features by using Wikidata to relate entities to countries, (ii) investigate the impact of the reimplemented feature on popularity in UK and US, and (iii) build a country-specific prediction model with these reimplemented features.

The Twitter data collection and processing methods were approved by the Mathematics and Physical Sciences and Engineering Joint Faculty Research Ethics Committee (application reference: MEEC16-031).

We first describe how we inferred user country from Twitter data in Section 9.1. In Section 9.2 we outline how we used Wikidata to relate TagMe entities to specific countries. Then in Section 9.3 we outline and evaluate the reimplementation of Proximity and Prominence using this new data, and finally in Section 9.4 we report the results of the country-specific prediction model.

9.1 Geolocationg Users on Twitter

Our goal is to obtain geographic location of users whose tweets contain a URL to an article in our datasets. Then we use the number of tweets geolocated to a specific country as a popularity measure for the country-specific prediction model.

Overview of Twitter geolocation methods. Approaches to geolocation on Twitter can be categorised along multiple axes. Firstly, according to geolocation target – tweets or users. Secondly, according to geolocation approach – content-based or network-based. Thirdly, according to granularity – from exact GPS co-ordinates to country or continent level.

The first two aspects (tweets or users, content- or network-based approach) are linked. Approaches which aim to infer user location commonly make use of the social network (e.g. check the user's friends' locations). Correspondingly, approaches that aim to geolocate a tweet (i.e. where was the message sent) commonly use tweet content (both message text and tweet metadata). Both content- and network-based approaches have attracted considerable research attention (cf. Jurgens et al. (2015) for an overview of network-based approaches; Zubiaga et al. (2017) provides a brief summary of both content- and network-based approaches). Geolocation granularity tends to vary as well, however city-level granularity seems to be most common.

Rout et al. (2013) and Compton et al. (2014) used network-based approaches which geolocate Twitter users. Both make use of a given user's social network graph to infer their location. Rout et al. (2013) geolocates to city-level, while Compton et al. (2014) geolocated to an exact location (GPS coordinates). Eisenstein et al. (2010) and Graham et al. (2014) used content-based approaches which geolocate tweets. Using tweet content Eisenstein et al. (2010) obtained topics which are specific to certain regions, while Graham et al. (2014) used the user-provided location field. The considerable advantage of content-based methods is that they do no require any further data collection apart from the tweets in the original corpus. On the other hand, network-based approaches require mining of the social network for all users in the dataset. They are also more computationally intensive.

Following the above, we chose a content-based approach. We apply geolocation to tweets, since that is the content of our dataset. Finally, we chose a coarser granularity of a country-level geolocation, instead of more fine-grained locations like region or city, to ensure a greater coverage of geolocation for the datasets. Country-level location is sufficient for our purposes (investigating Prominence and Proximity at country level). In the future, finer-grained investigations can be conducted.

9.1.1 Method for Country-Level Geolocation on Twitter

We use the Carmen software (Dredze et al., 2013). It is one of the few methods that provides worldwide tweet geolocation at a country level based on tweet content (cf. Zubiaga et al. (2017)). The system was evaluated on over 56,000 geolocated tweets and reached at least 90% accuracy at the country level and drops to above 50% at lower levels of granularity (Dredze et al., 2013). As our target granularity is country-level, we were satisfied with the system's performance.

Carmen aims to provide a tweet location from a database of structured location information. The database consists of locations with a set of coordinates and location names at different granularities (city, county, state, country). Null values in names are allowed to support underspecification (e.g. only country-level geolocation for *United Kingdom*). The system uses three methods of resolving location in a tweet:

"Place" object: structured field provided by Twitter in some tweets. This is queried against the location database.

Latitude and longitude coordinates: based on user's GPS position. This is queried against any database location within 25 miles.

User profile: free-form location field. Normalised strings are matched against the database.

9.1.2 Application on Headlines Corpora

We ran the Carmen system on the tweets in our dataset. The default setting in Carmen has the following order of resolvers: "Place" object, GPS coordinates, user profile. However, for our task user profile is the most appropriate, because people tend to put their city/country of residence in their profiles ("Place" object and coordinates might incorrectly geolocate

Table 9.1: Coverage of geolocation across datasets at different granularity levels.

	# articles	# tweets	# geolocated tweets (%)			
			city	state	country	total
The Guardian training	11,980	2,289,788	641,669 28%	99,865 4%	295,618 12%	1,063,642 46.45%
The Guardian test	13,806	2,244,885	623,317 28%	93,766 4%	346,206 15%	1,085,736 48.37%
New York Times training	5074	1,885,901	529,249 28%	167,076 9%	139,155 7%	838,081 44.44%
New York Times test	5011	2,206,585	584,214 27%	165,119 8%	166,027 8%	918,389 41.62%

users when they travel). Order of resolvers in our experiments was set to: user profile, "Place" object, GPS coordinates.

Table 9.1 presents the coverage of our chosen geolocation method in our datasets. In all cases at least 40% of the tweets in the datasets are geolocated to at least a country level. The majority of geolocated tweets (corresponding to at least 27% of all tweets in the datasets) are geolocated to a much finer granularity of city level. While our focus is on country-level location of users, when reimplementing Prominence and Proximity we infer country relatedness from lower level of location granularity (e.g. city) using the Wikidata knowledge graph, which we describe in Section 9.2.

While the geolocation method covers nearly half of all tweets in the datasets, we found that the coverage across articles is much better (cf. Table 9.2), where across datasets 87% of articles have at least one geolocated tweet. However, the number of geolocated tweets varies per dataset, with the median number of geolocated tweets much lower for *The Guardian* (median=16) than *New York Times* (median=37). This might be due to how common geotagging with GPS or supplying a location is across different user demographics on Twitter. In order to ensure the validity of our country-specific popularity, we compared the overall and geolocated median popularity and found no significant differences.

Next we look at how geolocated tweets are distributed across countries (cf. Table 9.3). We do this to establish for which countries we have a sufficient number of tweets per headline, so that we can build a prediction model in Section 9.4. We focused on countries where the median number of geolocated tweets is greater than zero. For *The Guardian* there are three such countries, while for *New York Times* there are seven. The disparity can be explained by the higher overall number of tweets in *New York Times* datasets.

Table 9.2: Summary of geolocation across articles in training datasets.

	The Guardian	*New York Times*
Number of articles	11980	5074
Number of articles with ≥ 1 geolocated tweet (%)	11401 (95%)	4431 (87%)
Median number of located tweets per article	16	37
Number of countries	171	176
Overall tweets median popularity (T1)	38	103
Geolocated tweets median popularity (T1)	40	108

Table 9.3: Summary of country-level geolocation in our datasets.

Dataset	0 <median countries	Median # geolocated tweets	# articles with 0<geolocated tweet (% total)
The Guardian	UK	6	10,682 (89%)
	USA	2	9,241 (77%)
	Canada	1	6,908 (58%)
New York Times	USA	29	4,414 (87%)
	UK	1	2,565 (51%)
	India	2	3,533 (70%)
	Canada	1	2,834 (56%)
	Mexico	1	3,218 (63%)
	Singapore	1	3,189 (63%)
	Italy	1	3,273 (65%)

Median number of geolocated tweets is generally low. Highest medians are recorded for the country where a given news source is based: median=6 for UK in *The Guardian* dataset, and median=29 for US in *New York Times* dataset. When looking at the number of articles in a dataset with at least one geolocated tweet, the coverage is also the best for news sources' country of origin: 89% of articles in *The Guardian* training dataset have at least one tweet geolocated to UK, and 87% of articles in *New York Times* training dataset have at least one tweet geolocated to US. In order to ensure sufficient differentiation in popularity we limit country-specific models to those where the median number of geolocated tweets is highest, i.e. popularity in United Kingdom for *The Guardian* corpus, and popularity in the United States for *New York Times* corpus. A focus on UK-based users for *The Guardian* and US-based users for *New York Times* also follows the results of our qualitative analysis in Section 7.3. In that analysis we noted that familiarity with the news source has a bearing on how a headline is interpreted. Residing in the country of the news source can lead to a higher familiarity with it.

Table 9.4: Number of cases where home country popularity is lower than popularity in another country. Reported are *All Cases* (all cases regardless of popularity), *0 <Home country popularity* (cases where home popularity is higher than zero).

Dataset	All Cases	0 <Home country popularity (%)
The Guardian	6113	3976
New York Times	46	23

Table 9.5: *The Guardian* headlines where non-UK popularity was higher than UK popularity.

Headline	Popularity at home	Popularity abroad
"New York assault weapons ban circumvented with simple modification"	UK: 12	US: 63
"Australian government may ban environmental boycotts"	UK: 58	Australia: 305
"Move over Heathrow. Now Dubai International is the world's No 1 airport"	UK: 33	UAE: 63
"Narendra Modi as prime minister would roll back women's rights in India"	UK: 10	India: 75
"Silvio Berlusconi's judgment day – you choose the sentence"	UK: 30	Italy: 76

9.1.3 Initial Insights Using Country-Specific Popularity

As an initial investigation into whether geographic proximity influences news article popularity, we looked at cases where the popularity in the news outlet country (hence, home country) was lower than in some other country. For example, a case where an article in *New York Times* was tweeted 20 times in US, but 200 in India. Such cases might point to examples of articles which mention an entity relevant to another country, which leads to higher popularity in that country. Table 9.4 presents the number of such cases in each dataset. Cases where Twitter popularity is higher abroad than in the news outlet country are much more common for *The Guardian*. Factors like global and national readership structure might influence these results. We looked at a sample of these headlines to try and find an explanation for the higher Twitter popularity abroad. Tables 9.5 and 9.6 present examples of such headlines.

In the examples given in Tables 9.5 and 9.6 we observe that an entity relevant to another country was mentioned in the headline. These examples show us two things. Firstly, even though these events would be reported chiefly in the national media of the relevant country, readers from those countries still chose to share articles on these national topics using

Table 9.6: *New York Times* headlines where non-US popularity was higher than US popularity.

Headline	Popularity at home	Popularity abroad
"In Calgary, Exploring the Cultural Side of 'Cowtown' "	US: 33	Canada: 412
"Narendra Modi's American Facebook Fans"	US: 95	India: 158
"Australian Premier Moves Swiftly Against ISIS, but Analysts Question Benefits"	US: 27	Australia: 79
"New Leader Takes Oath of Office in Indonesia"	US: 37	Indonesia: 52

The Guardian or *New York Times* content. In the global community context, this might mean that sharing a news article in English from a major news outler might gain a wider response than sharing an article from a national (potentially non-English) news source. Secondly, named entities in headlines act as a signal of geographic relevance for readers. The entities are not just explicit mentions of other countries (e.g. *India, Indonesia*), but also people (e.g. *Narendra Modi*, *Silvio Berlusconi*), cities (e.g. *New York*, *Calgary*), and other types of locations (e.g. *Heathrow*). Since these different types of entities can indicate relatedness to a country, it is crucial that any type of entity (and not just location) is considered for Proximity and country-aware Prominence. In order to link these different types of entities to a country, we need to use world knowledge. We do this by utlising the Wikidata knowledge graph (cf. Table 9.7). This Table shows examples of direct connections of a variety of entities to several country nodes. However, in our experiments we need to account for more than direct connections.

9.2 Relating Entities to Countries

Our goal is to establish whether a given TagMe entity is related to a country (United Kingdom or United States). As we have to consider a variety of entities and thus a variety of potential relations between the entities and countries, we make use of the Wikidata knowledge graph. Since we are using TagMe which outputs a Wikipedia page for each entity, Wikidata is the natural knowledge graph to use. Our method of relating TagMe entities to countries is carried out two steps: firstly, we find the Wikidata ID corresponding to the TagMe concept, and secondly, we establish whether a given Wikidata ID (i.e. an entity) is closely connected to the UK/US node in the Wikidata graph.

Table 9.7: Examples of entities in Tables 9.5 and 9.6 and their relation to a country in Wikidata. Length of path refers to the number of edges necessary to connect the nodes. Relation refers to the property that connects the nodes.

Text	TagMe entity	Related country	Length of path	Relation
New York	NEW YORK	United States of America	1	*country*
Dubai	DUBAI INTERNATIONAL AIRPORT	United Arab Emirates	1	*country*
Silvio Berlusconi	SILVIO BERLUSCONI	Italy	1	*country of citizenship*
Calgary	CALGARY	Canada	1	*country*
Narendra Modi	NARENDRA MODI	India	1	*country of citizenship*

Wikidata. We use an RDF export of Wikidata which is closest in time to our datasets (20th April 2014). The dataset is available from Wikimedia Tool Labs[1]. The RDF exports are split into:

- *Terms*: labels, descriptions, and aliases in all languages
- *Properties*: property definitions, including datatypes, labels, descriptions, and aliases
- *Statements*: simplified statements (one triple per statement) without references or qualifiers

The exports were loaded into GraphDB[2], which was then used to run SPARQL queries. The SPARQL queries we used are included in Appendix G.

9.2.1 Obtaining Wikidata IDs for TagMe Tags

Before we relate an entity to a country we first need to find the corresponding Wikidata ID (of the format *Q...*) for all TagMe entities in the combined *The Guardian* and *New York Times* corpora. There are 13,676 unique TagMe tags in the combined corpus. We ran a SPARQL query against the Wikidata Terms triplestore. The query (cf. Listing G.1 in Appendix G) searches only the English language terms and limits the results to the top query. Out of the total of 13,676 TagMe tags we were able to obtain Wikidata IDs for 10,891 (80%).

[1] http://tools.wmflabs.org/wikidata-exports/rdf/exports/20140420/dump_download.html [Accessed 13th April 2018]

[2] http://graphdb.ontotext.com/ [Accessed 13th April 2018]

Table 9.8: Counts and percentages of Wikidata IDs related by paths of lengths [1,3] to UK or US Wikidata nodes.

	UK		**US**	
	Count	Percentage	Count	Percentage
Length=1	790	7%	811	7%
Length=2	1043	10%	777	7%
Length=3	2893	27%	2396	22%

Total number of Wikidata IDs: 10,891

9.2.2 Connecting Tags to the UK/US Node in Wikidata Graph

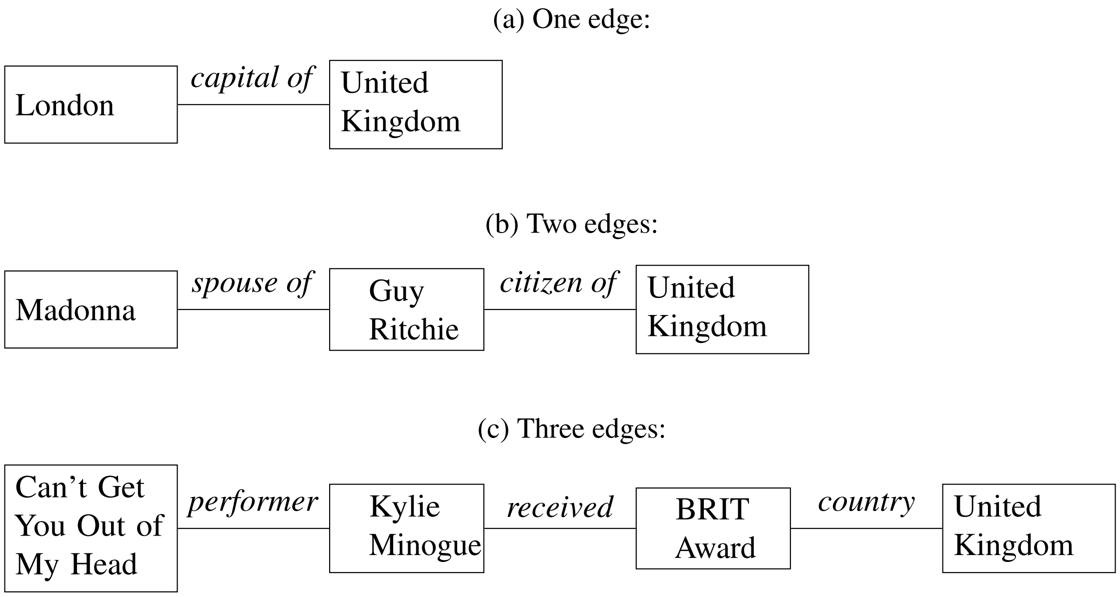
Our next step was to find whether a given Wikidata node (which corresponds to a TagMe entity) is closely related to either United Kingdom (in case of *The Guardian*) or United State (in case of *New York Times*). The relation between an entity node and UK/US node is captured by the path in the graph between the nodes. Since we are interested in whether a given entity is closely related to UK/US, we limit the path length to between one (i.e. direct connection) and three edges. Listings G.2 to G.6 in Appendix G show the queries for relating entities to the two country nodes using the example of the node for entity *United Kingdom* (Wikidata ID: Q145). A representative example of a query is included in Listing 9.1. The query asks if there are any properties (`?p`) which connect a given Tagme entity identified with a Wikidata ID (`WIKI_ID`) to the United Kingdom node in Wikidata (`wd:Q145`).

```
PREFIX wd: <http://www.wikidata.org/entity/>
ASK
WHERE {
        WIKI_ID ?p wd:Q145 .
}
```

Listing 9.1: SPARQL query for connection to UK with one edge

Table 9.8 presents the results of running the queries for this dataset and examples of relations of the three path lengths are presented in Figure 9.1. Within the path length of [1,3] there were roughly similar counts of related Wikidata IDs for both UK and US, with the counts for UK being slightly higher. Approximately 7% of Wikidata IDs were directly connected to either UK or US node; between 7% and 10% had two edges, and between 22% and 27% had three edges. This means that over a third of Wikidata IDs in our datasets (UK: 43%, US:37%) had a relatively close connection to the two country nodes using these

Figure 9.1: Examples of entities related to United Kingdom with different path lengths.



(a) One edge:

(b) Two edges:

(c) Three edges:

paths.

9.2.3 Evaluation

In order to validate our method of relating TagMe entities in headlines to countries, we compared the output of our automatic method against a manually annotated gold standard. Our goal is to establish that the automatic method correctly identifies *if* an entity is related to a country, and correctly estimates the *strength* of that relation.

Obtaining the gold standard. We randomly sampled 100 entities we have in our dataset. Because we are working with headlines corpora from two countries, we created two gold standard datasets – one of United Kingdom (to be used with *The Guardian* headlines) and one for United States (to be used with *New York Times* headlines). The same set of 100 entities was used for both gold standards, since the random sample already includes some UK-related and US-related entities, as well as entities which are not related to either. We had three annotators working on each gold standard dataset. The annotators were British/American citizens living in United Kingdom/United States. Apart from the country of residence, there were no restrictions placed on annotator eligibility for this task.

The annotators were asked to score each of the 100 entities on a 0-3 Likert scale depending on the strength of the relation between a given entity and UK/US. The scale item descriptors are presented in Table 9.9. Because lack of familiarity with a given entity will not allow an annotator to assign it a score, we asked the annotator to score any unfamiliar entity with zero and indicate the entity as 'not familiar'. The entities in the

Table 9.9: Scale used for assigning relatedness between an entity and UK/US in the manual annotation task.

Scale item	Description
0	not related at all
1	slightly related
2	considerably related
3	very closely related

Table 9.10: Fleiss's kappa inter-annotator agreement for UK and US gold standard for entity-country relatedness.

Setting	UK gold standard	US gold standard
Strict (0-3 scale)	0.39 (fair)	0.35 (fair)
Relaxed (binary)	0.48 (moderate)	0.54 (moderate)

random sample that *none* of the annotators were familiar with were: SPIDEROAK, KERRY STOKES, JOSÉ SARAMANGO, JAZMIN CARLIN, DARIUS BOYD (British annotators); DYKES TO WATCH OUT FOR, ANT & DEC (American annotators).

We next calculated the inter-annotator agreement using Fleiss' kappa (Fleiss, 1971). We chose Fleiss' kappa because our data was ordinal and we had more than two annotators. Results are reported in 9.10. We used two settings: strict (considers the whole 0-3 scale) and relaxed (binary; any score over and including 1 was set as 1). The relaxed setting gives an indication whether there is any relation at all between an entity and UK/US perceived by annotators, regardless of relation strength. Taking into account the varying levels of familiarity, the agreement between annotators for both gold standard datasets is acceptable. These levels of inter-annotator agreement highlight that even for humans this is a difficult task. The annotators' judgments were aggregated using a majority vote, yielding a gold standard for comparison with the automatic method.

Next we compare the results of the automatic method with the gold standards. We treat this as a classification task using an exact and relaxed accuracy score. The exact accuracy considers the full 0-3 scale, whereas the relaxed accuracy takes into account the binary score (related vs. unrelated). We report the accuracy scores in Table 9.11. With both strict and relaxed accuracy we achieved approximately 70% accuracy. Considering the challenging nature of this task (e.g. implicit vs. explicit relatedness we mentioned), we are satisfied with these results. Our next step was to use the Wikidata path lengths to augment Prominence and Proximity features from the global popularity model. This allowed us to

Table 9.11: Accuracy results for entity-country relatedness which compare the automatic method against the human gold standard.

Setting	UK gold standard	US gold standard
Strict (0-3 scale)	69%	70%
Relaxed (binary)	71%	72%

investigate our hypothesis that geographic relevance impacts the social media popularity of news articles.

9.3 Tuning News Values

In the previous section we presented our methods for relating entities in headlines to countries (UK and US) using Wikidata. In this section we outline how these entity-country relations are used to reimplement the news values of Prominence and Proximity.

9.3.1 Boosting

The main change of this reimplementation is integrating the additional information about entity relations to UK/US to the feature engineering method. In literature there are many examples of using semantic resources like Wikidata to enrich NLP representations (e.g. Gabrilovich and Markovitch (2009)), or to generate new features using Linked Open Data (Paulheim and Fümkranz, 2012). Ristoski and Paulheim (2016) included an overview of common methods for feature generation using semantic data. These include: binary feature indicating whether a relation between concepts exists, numeric feature counting the number of relations of a certain type, and using the literals from the triples as nominal or numeric feature values.

However, none of these methods corresponds exactly to our task, i.e. tuning an existing feature implementation with semantic information contained in a limited number of discrete values. As an initial investigation into feature tuning using semantic data, we propose boosting. Our only requirement is that the boosting is proportionate to the path length on Wikidata. Since we only have three path lengths (path lengths=[1,3]), we can create three boosting rules (one for each path length). In the case of Prominence since we are boosting existing values (i.e. the entities' Prominence), we multiply the feature value by a boosting weight (cf. Tables 9.12 and 9.13). In the case of Proximity since we currently have a binary feature, we replace it with the path length (cf. Tables 9.14 and 9.15).

9.3.2 Prominence

The prominence score of an entity that is geographically relevant to a reader should be proportionate to the relevance of that entity. We reimplement Prominence by utilising the path length between an entity and the UK/US node to boost the Wikipedia prominence score.

Our reimplementation rests on a simple rule: if there are no entities that have geographic relevance to the reader, then leave the prominence scores at their current values; else if there is at least one entity that has geographic relevance to the reader, then that entity's prominence score is boosted in proportion to the relevance to the reader's country as measured by the Wikidata path length. All the other details of Prominence implementation remain the same: entities' Wikipedia prominence is measured using pageviews (long-term prominence for entity *e*: $pageviews_{e,d_{-365},d_{-1}}$, day-before prominence: $pageviews_{e,d_{-1},d_{-1}}$) and entities' prominence in a headline is still aggregated using the sum.

We experimented with several different versions of the boosting, optimising for correlation with Twitter popularity (cf. Tables 9.12 and 9.13).

Table 9.12: Experiments with different boosting methods for Prominence (*The Guardian* training set). *i* denotes an entity's prominence. Dist... refers to distances between entities in the graph and the corresponding boosting for an entity. τ refers to Kendall's correlation with UK popularity.

	Dist>3	Dist=1	Dist=2	Dist=3	τ (long-term)	τ (day-before)
No boosting	i	i	i	i	0.0495	0.0428
$Boosting_1$	i	$i * 2$	$i * 1.6$	$i * 1.3$	**0.0507**	**0.0441**
$Boosting_2$	i	$i * 10$	$i * 6$	$i * 3$	0.0459	0.041
$Boosting_3$	i	$i * 100$	$i * 60$	$i * 30$	0.0291	0.0264

all correlations are significant at p<0.001

Our exploration of boosting using discrete semantic data yielded mixed results. We saw a slight improvement in the correlation for two Prominence features in *The Guardian* dataset. However, the same boosting methods in *New York Times* dataset showed no improvement. Greater path lengths or a fine-grained parameter search might find a boosting method which does improve the correlations. The positive results for *The Guardian* encourages further investigations.

9.3.3 Proximity

Proximity has been implemented as a binary feature in the global popularity model. We reimplemented it using the path lengths that relate entities to countries using similar

Table 9.13: Experiments with different boosting methods for Prominence (*New York Times* training set). *i* denotes an entity's prominence. Dist... refers to distances between entities in the graph and the corresponding boosting for an entity. τ refers to Kendall's correlation with US popularity.

	Dist>3	Dist=1	Dist=2	Dist=3	τ (long-term)	τ (day-before)
No boosting	i	i	i	i	**0.0344**	**0.0398**
$Boosting_1$	i	$i * 2$	$i * 1.6$	$i * 1.3$	0.0342	0.0397
$Boosting_2$	i	$i * 10$	$i * 6$	$i * 3$	0.0338	0.0392
$Boosting_3$	i	$i * 100$	$i * 60$	$i * 30$	0.0335	0.0377

all correlations are significant at p<0.01

boosting methods to the ones we used for Prominence. Instead of boosting an existing value (since the existing values are binary), we replace it with a value proportional to the path length. Results are presented in Tables 9.14 and 9.15.

Table 9.14: Experiments with different boosting methods for Proximity (*The Guardian* training set). *i* denotes an entity's prominence. Dist... refers to distances between entities in the graph and the corresponding boosting for an entity. τ refers to Kendall's correlation with UK popularity.

	Dist>3	Dist=1	Dist=2	Dist=3	τ	Sig.
No boosting	i	i	i	i	N/A	N/A
$Boosting_1$	0	3	2	1	0.0083	0.24
$Boosting_2$	0	10	6	3	**0.0094**	**0.18**
$Boosting_3$	0	100	60	30	**0.0094**	**0.18**

Conversely to the tuning results for Prominence, in the case of Proximity the tuning seems to only work for *New York Times* dataset. Since the feature is now numeric instead of binary, we cannot make a direct comparison with the original implementation. However, the correlation with US popularity is statistically significant. That is not the case for *The Guardian* dataset where the reimplemented Proximity feature does not correlate with UK popularity at a statistically significant level. We would like to highlight that the correlation in *New York Times* dataset is positive, that is to say the more closely related entities are to US, the more popular the headline. This is opposite to the effect that the original Proximity feature had on global popularity (Tables 6.1 and 6.2). The new feature implementation method which includes user country information allows for a more accurate investigation of this news value. This new finding follows journalistic literature which states that geographic proximity should have a positive influence on newsworthiness.

The further investigation into tuning Prominence and Proximity features using semantic

Table 9.15: Experiments with different boosting methods for Proximity (*New York Times* training set). *i* denotes an entity's prominence. Dist… refers to distances between entities in the graph and the corresponding boosting for an entity. τ refers to Kendall's correlation with US popularity.

	Dist>3	Dist=1	Dist=2	Dist=3	τ	Sig.
No boosting	i	i	i	i	N/A	N/A
$Boosting_1$	0	3	2	1	**0.0276**	**0.02**
$Boosting_2$	0	10	6	3	**0.0276**	**0.02**
$Boosting_3$	0	100	60	30	**0.0276**	**0.02**

information (i.e. user location) yielded mixed results. We did see a slight improvement in correlations for Prominence features in *The Guardian* and a significant correlation for Proximity in *New York Times*. Crucially, with feature implementation that is aware of user location we were able to clarify our findings on Proximity, at least for *New York Times* corpus.

9.4 Country-Specific Prediction Model

In the previous section we presented the reimplementations of Wikipedia-related Prominence and Proximity features. We showed that retuning the features with entity relatedness information from Wikidata improved feature correlations compared to location-agnostic versions in all cases except the Prominence reimplementation for *New York Times*. Next we include these reimplemented features in the country-specific popularity prediction model, which we then compare to the standard model (i.e. with original feature implementations). Our goal is to establish whether knowledge-enhanced features which reflect geographic relevance improve prediction results.

Because the correlation results did not clearly point to one boosting method achieving significantly higher results we chose the following settings aiming to maximise the prediction performance. For both datasets we use $Boosting_1$ for long-term and day-before Wikipedia prominence, and $Boosting_2$ for Proximity. The prediction results are reported in Table 9.16. We found no significant differences between the prediction models which used features reimplemented using location information or not. Since we did observe some slight improvements in the impact of the reimplemented features on Twitter popularity, we are encouraged to continue working on better reimplementations.

Table 9.16: Comparison of prediction results between the location-aware and location-agnostic models.

	The Guardian		*New York Times*	
	τ	MAE	τ	MAE
Location-agnostic	0.35	0.76	0.1	0.72
Location-aware	0.35	0.76	0.09	0.72

9.5 Summary

Our findings in earlier chapters indicated the need for a user location-aware investigation of Prominence and Proximity. In this chapter we created a corpus of headlines with country-specific Twitter popularity. This provided us with evidence that a variety of entities can indicate relatedness to a country (cf. Section 9.1.3). That initial investigation pointed to the necessity of using knowledge graphs in order to be able to connect a variety of entity types with countries. This led us to use Wikidata in order to connect TagMe entities identified in headlines to the United Kingdom node (for *The Guardian*) and United States node (*New York Times*) in order to establish their relatedness. We validated our methods against a manually annotated gold standard. With the additional country relatedness information we reimplemented Prominence and Proximity using boosting. We found that in all cases (except Prominence in *New York Times*) the reimplemented features correlated more highly with Twitter popularity. The prediction models using the reimplemented features did not significantly improve over location-agnostic models.

There are several factors which could have influenced our results. Firstly, by using a resource like Wikidata we are dependent on explicitly stated relationships between entities. In case of people and locations the relationship to a country is usually clearly stated (e.g. London → *capital of* → United Kingdom). However, for some concepts the relationships are more implicit. For example, among the TagMe tags in our dataset we have an entity *Tea* (referring to the beverage). Using the Wikidata graph we found no connection within our span between the *Tea* node and the *United Kingdom* node, although most people would readily associate the two concepts. Longer paths and restricting properties to only certain types might further improve the country relatedness for entities. Secondly, there might be other boosting methods or parameters to be explored. We showed that our simple method already significantly improves over location-agnostic feature implementations. This presents a strong case for the need to include world knowledge into feature engineering and considering user characteristics for news article popularity prediction tasks.

Our main contributions for this chapter are:

(i) we developed and validated a method for relating headline entities to countries using Wikidata

(ii) we proposed an approach to augmenting linguistic features with world knowledge using simple boosting

(iii) we showed that location-aware implementation of Prominence and Proximity correlate better with Twitter popularity than location-agnostic implementations.

Chapter 10

Conclusions

In this thesis we present our approach and findings on using headlines to model the social media popularity of news articles. We followed an experimental methodology to identify and operationalise explicit linguistic indicators in the headline text which influence news articles' popularity on social media. We investigated two types of features: journalism-inspired news values and linguistic style. We tested the impact of the proposed features in a number of experiments. First, we correlated feature values with social media popularity, measured by the number of reactions on Twitter and Facebook. We also conducted a quantitative and qualitative analysis of the impact of news values and style features on perceived popularity, obtained from a crowdsourced survey. This gave us an indication of the impact of individual headline features and the need for including user demographics in feature implementations. To investigate the predictive power of all features combined, we trained and evaluated a number of prediction models, taking into account different feature and corpus subsets. Finally, using Twitter data we obtained a country-specific Twitter popularity measure which was then used to evaluate features augmented with country information from a knowledge graph. Our main contributions are the operationalisation and evaluation of news values features in headlines, thorough investigation into the impact of headline-derived news values and style features on two types of popularity measures, and the training and evaluation of global and country-specific social media popularity prediction models using headlines.

In this chapter we first provide a synopsis of this thesis. We then reflect on the methodology we used and its impact on findings. Next, we summarise our contributions to relevant research fields. Finally, we provide suggestions for future work.

10.1 Synopsis

In this thesis we modelled the social media popularity of news articles using headline text. For that purpose we created two headlines corpora using data from *The Guardian* and *New York Times*, and obtained the associated news article popularity on Twitter and Facebook.

In Section 1.3 we posed five research questions. In the following paragraphs we list these research questions and summarise our findings for each one.

RQ1: *Can news values be reliably extracted from headline text?*

For the task of modelling the social media popularity of news articles using headline text, we proposed using news values – newsworthiness factors described in journalism studies. In Chapter 4 we presented the development and evaluation of computational methods for extracting six news values from headline text: Prominence, Sentiment, Magnitude, Proximity, Surprise, and Uniqueness. For implementing Prominence we introduced the use of wikification and burstiness. For Sentiment, we used direct (sentiment, polarity) and proposed the use of indirect measures (connotations, biased language). For Magnitude, we used topic-independent indicators in the form of comparatives/superlatives, intensifiers, and downtoners. For Proximity, we used a wordlist of UK/US-related terms and looked for matches in the headline text and Wikipedia categories of entities that occur. For Surprise, we focus on surprising phrasing measured by the commonness of syntactic chunks in relation to a large corpus. For Uniqueness, we compared headlines using cosine similarity, but considered whether there is also an entity overlap between them. In a comparison with a gold standard we found that for each proposed news value we can reliably identify it in headline text. Some proposed feature extraction methods (e.g. some aspects of Sentiment) might benefit from further development. We also identified the need for a user-aware feature extraction method for some news values (Prominence, Proximity), which led us to build a country-specific model discussed later.

RQ2: *What is the impact of headline-derived news values and style features on social media popularity?*

In Chapter 6 we investigated the impact of both news values and linguistic style features on news article popularity on Twitter and Facebook. We found that for each news value and style aspect at least one feature had a statistically significant impact on social media popularity. We noted differences between the features' impact between news values and style features, between the two news sources, and between Twitter and Facebook. We found that most news values and verb-related style features were more strongly correlated with social media popularity in *The Guardian* corpus; whereas for *New York Times* it was Sentiment, Brevity and syntactic Simplicity features that had highest correlations. In terms of popularity measures, *The Guardian* correlations were generally higher for Twitter compared to Facebook, however in *New York Times* corpus Facebook usually had higher correlations. Some unexpected results (Proximity) indicated the need for a model aware of user location.

RQ3: *What is the impact of headline-derived news values and style features on perceived popularity and how is it judged by readers?*

In Chapter 7 we investigated perceived popularity – whether readers think a headline would be clicked on, which provides complementary insights to social media popularity (whether a headline would be shared). In order to obtain the perceived popularity measure, we conducted a survey using a crowdsourcing platform. We correlated the responses with feature values and found that a much smaller number had a significant impact, however in cases where the impact was significant the effect was greater than for social media popularity (e.g. Brevity and syntactic Simplicity features in *New York Times* corpus). We also conducted a qualitative analysis of survey responses by three experts and identified familiarity, genre, and domain specificity as factors influencing readers' decisions. The survey participants' judgements about news values and style revealed their perceived importance and how they differed from our findings about their impact on social media popularity, namely the higher rating of news values over linguistic style features.

RQ4: *To what extent can headline-derived news values and style features be used to predict the social media popularity of news articles?*

Having shown that individual news values and style features have a significant impact on social media popularity and perceived popularity, we then combined all features in prediction models of social media popularity in Chapter 8. We reimplemented state-of-the-art baselines by Bandari et al. (2012) and Arapakis et al. (2014) (which added pre-publication metadata to the model) and showed improvement over the baselines for all measures with the exception of MAE results for Facebook in *New York Times* corpus. We also ran the prediction models using feature and corpus subsets where we found that headline-derived features (news values and style) performed as well as metadata in *New York Times* and that controlling the news category can decrease the prediction errors.

RQ5: *Does augmenting the feature engineering with country-specific information improve the impact of that feature on social media popularity?*

In Chapter 9 we utilised user location data obtained from Twitter metadata in order to build a country-specific popularity prediction model. This allowed us to control for familiarity factors by taking into account reader location and to clarify our findings about the Proximity feature. As well as being able to predict popularity for users from a particular country, we also reimplemented Proximity and some Prominence features to take into account relatedness of headline content with a given country. For these reimplementations we utilised a novel approach to augment existing NLP feature engineering with world knowledge from the Wikidata knowledge graph. Including augmented features in a prediction model showed no significant improvement over a standard model, however our exploration revealed modest improvement over the location-agnostic feature extraction, which supports the combination of NLP and knowledge graphs for feature engineering.

10.2 Reflections on the Methodology

Our choice of methodology and the decisions we took throughout our experiments affect our findings.

Content beyond headline text. Our experiments are based on headline text. This has two implications: (i) other modes of discourse, such as image or layout, are not considered; and (ii) the production and reception context is limited. Regarding the first point, we acknowledge that non-textual elements such as images or layout influence social media popularity, and this effect is missed by our analysis. The second point – lack of analysis of production and reception context of news – has been taken up by Philo (2007); Carvalho (2008) who argued that the aspects of news outside of the text (e.g. interviews with news producers and news audiences and analysis of social actors) are essential for analysis. However, Fürsich (2009) presented a defence of text-only analysis, where she highlighted that textual analysis offers a wider set of possible readings which is more objective than the responses from producers and audience which necessarily limit the interpretations of news. Our experiments do not consider the production context with the small exception of using the style guide as an inspiration for some features. We do however consider the audience reception perspective by using the social media popularity measures. In particular, our quantitative and qualitative analysis using perceived popularity measures from a crowdsourced survey gives us insight into the audience response and what headline aspects might influence it.

Genres beyond broadsheets. We used two headlines corpora in our experiments (*The Guardian*and *New York Times*). Both were obtained from news outlets which are described as 'broadsheet' or 'quality' newspapers. This means that our proposed feature operationsalisations and our findings on the impact of these features are valid for the two news outlets we considered, as well as other news outlets of the same type, since broadsheets tend to display similar, even tone. When it comes to other news genres (e.g. tabloids) we envisage that our operationalisation methods will still be applicable, as our implementations are domain-independent. However, new features might be required (e.g. to quantify sensationalism) and the impact of our features might be different.

Use of existing NLP tools. Our feature operationalisations in Chapters 4 and 5 relied on the use of external NLP tools. In Section 2.3 we noted that headlines as a text type bring a number of challenges for computational processing. Ideally headline-specific NLP tools would be used for parsing and named entity recognition, however training or building such tools was beyond the resources of this project. In order to ensure reliability of results, the performance of third-party tools was evaluated on a small sample of headlines. The tools that were chosen achieved an acceptable level of accuracy. Due to the brevity of headline text, wherever possible we used tools (e.g. the use of TagMe for named entity recognition and entity linking), or methods (e.g. keyword-based methods for Sentiment) meant for short texts.

Generalisability of our approach. Our requirements for feature operationalisation (especially in the case of news values) was that the methods are domain-independent and applicable to other types of short text. Although we used corpora from the news domain, none of our proposed features rely on the text to be from the news domain. As the implementations rely on explicit and domain-independent methods, we envisage that our methods are applicable to any type of short text that is similar to headlines, for example, titles of other types of online content (videos or blog posts). Since we found our features to have significant impact on social media popularity, our methods could be of use to authors of tweets. To account for certain characteristics of tweets (e.g. usage of mentions and hashtags, emoticons, abbreviated text) Twitter-specific methods could be substituted for certain elements in our approach. For example, instead of TagMe a Twitter-specific entity linking tool can be used to obtain entities, however our novel proposals for entity burstiness and enriching NLP feature engineering with world knowledge by using Wikidata-based country linking can still be utilised.

Country-specific prediction. In Chapter 9 we used Twitter data to obtain country-specific popularity of the news articles in our datasets. Using the content-based geolocation method we were able geolocate nearly half of the tweets in our datasets. This resulted in very low median numbers of geolocated tweets for some countries, and led us to only consider home country popularity for the prediction task. Using a graph-based geolocation method (although more computationally intensive) might yield better coverage of the dataset. Furthermore, the location tagging behaviour of Twitter users might have changed since our data collection period, leading to a higher number of tweets with a location available through the Twitter API, and thus higher geolocation coverage. Since we only considered home country popularity, we cannot make a case that our results will generalise beyond the cases we considered. However, headlines with a higher popularity in a country related to an entity mentioned in that headline (Section 9.1.3) offer some promising insights.

10.3 Contributions

The main contributions of this thesis are in three domains.

10.3.1 Contributions to Natural Language Processing and Knowledge-Enriched Feature Engineering

One of our main contributions is the development and evaluation of computational methods for operationalising news values from headline text. We utilised a range of state-of-the-art NLP methods in order to reliably capture news values. One of our most innovative feature operationalisations is for the news value of Prominence (cf. Section 4.1.1). We proposed the use of Wikipedia-based entity linking (i.e. wikification), instead of standard named entity recognition. By linking entity mentions in text to Wikipedia concepts, we were able to access a semantic-rich resource that is Wikidata. This allowed us to develop feature implementations that considered relatedness of a given contept to a country, which was then utilised in the country-specific prediction model in Chapter 9. We also proposed the use of a burst detection algortithm in order to take into account the temporal variability of entity prominence. The operationalisation for the news value of Uniqueness also made use of Wikipedia entities. We found that checking for overlap in entities between two headlines was helpful in ensuring that they were part of the same storyline. This provided a deeper level of similarity. These methods can be applied to other popularity modelling tasks which make use of text-derived features.

10.3.2 Contributions to Research on Computational Methods in Web and Social Media

We made a number of contributions to the research on computational methods for the Web and Social Media. Firstly, we created and shared two datasets: (i) headlines corpora from *The Guardian* and *New York Times* annotated with news values (cf. Appendix A). Secondly, we evaluated the features we proposed by looking at the impact of individual features (correlations with social media popularity and quantitative and qualitative analysis of perceived popularity), as well as the impact of combining features in a global and country-specific prediction model. Our approach which looks at two types of popularity measures (one obtained from social media, and one obtained from a crowdsourced study) can be replicated for other popularity modelling tasks. Using both types of popularity provides a 'big' and 'small' data perspective on the research problem.

10.3.3 Contributions to Digital Humanities

Our main contributions to the field of digital humanities are: (i) operationalisation and validation of news values extraction from headline text, and (ii) investigation into the impact of headline-derived news values and linguistic style on popularity. The operationalisation of news values will enable digital humanities researchers working with news corpora to automatically extract news values scores from headlines at scale. This in turn can enable the comparison of news values across different news outlets, genres, or demographics. Our in-depth investigation of the impact of news values and linguistic style of headline can inform the development of new versions of news values taxonomies and update guidelines on headline writing.

10.4 Future Work

We see several research directions for future work in the short-term and in the long-term.

10.4.1 Short-Term Improvements

We see two immediate improvements to our current work. Firstly, the study design for the crowdsourced study could be further refined. In particular, controlling for the participants' familiarity with the news source, as well as knowledge and interest of topics could clarify some of our findings. Furthermore, the method for obtaining the perceived popularity measure could be changed by asking the participants for preference between a pair of headlines, instead of giving a rating on a five-point scale. Secondly, other methods for augmenting Prominence and Proximity features with country relatedness information can be explored. A more sophisticated weighting which takes into account the relation type that links entities (e.g. *president of* being more relevant than *citizen of*) might produce better results.

10.4.2 Long-Term Directions

There are several research directions that can be investigated.

Generalising to other corpora. In this work we have established that news values and style features in headlines have a statistically significant effect on *The Guardian* and *New York Times* news article popularity on social media. Further studies are needed to establish to what extent these features and their impact generalise to other headlines corpora. We see three main research directions:

(i) headlines in other 'broadsheet' newspapers

(ii) headlines in other newspaper genres such as tabloids, or technical and scientific news outlets

(iii) other types of texts similar to headlines, such as video or blog post titles

We believe that in the case of other 'broadsheet' newspapers our methods can be applied without any changes. As we have already observed some differences in feature impact between *The Guardian* and *New York Times*, we expect that there will be some variety in terms of feature impact for other 'broadsheets' as well. However, at least commonality in the direction of the correlation (positive or negative) between different news outlets would suggest a general pattern. In case of other news genres, such as tabloids or scientific press, our methods could be applied without any modification, however it might be beneficial to consider adding some genre-specific features, such as sensationalism for tabloids, and number of academic citations for scientific press. Similarly, our methods could be applied to other types of short texts which function similarly to headlines, e.g. titles. However, both in case of other news genres and other text types we cannot make a prediction as to which features will have the most impact on popularity. Taking into account all these different text types, it would be interesting to see whether there are any overarching patterns, e.g. that Prominence always positively influences popularity on social media, or are there any cases where the impact is actually negative. These broad patterns would then be able to inform research studies in digital humanities about audience perception and consumption of online content.

Comparison of news values across corpora. Our proposed features implementations for news values and linguistic style in headlines can become a useful tool for discourse analysts. Our methods could be applied to large datasets of news headlines corpora, which is currently impossible when manually annotating news values. For example, it would be possible to explore the usage of news values across different news outlets or investigate temporal patterns of use within the same news outlet. Since the headlines corpora we created contain metadata about article categories, the corpora can be split into subset and used to compare news values and linguistic style across topics and genres. Our method could be complementary to the in-depth qualitative analysis and provide some new insights in digital journalism.

Hybrid approaches to news article popularity prediction. We showed a statistically significant effect of features derived from headline text on social media popularity. Since headline text is not the only factor which influences social media popularity of a news article (visual presentation and social effects can also be considered), the features we proposed in this thesis can be used together with features which implement those other factors. The advantage of using headline-derived features is that they are available for a 'cold start' problem in recommendation system and for pre-publication prediction systems. The interaction and predictive power of content (headline-derived) and context (visual, social) features should be explored, in order to build prediction systems which can use different features depending on when the prediction is made, and what data is available at that time point.

Knowledge-enriched feature engineering. The reimplementations of Proximity and some Prominence features which take into account user country showed some modest improvements over location-agnostic implementations. We have shown that knowledge-enriched feature engineering achieves very good results (e.g. Promience). We have also noted that our focus in this work is on explicit linguistic indicators of news values. However, operationalisation of news values could benefit from knowledge-enriched feature engineering. Sentiment features could take into account the types of entities that occur in a headline (e.g. disease-related entities like *Ebola* could indicate negative sentiment). Magnitude could be operationalised to take into account world knowledge, by contextualising any mentions of numbers (e.g. five is a high number of victims in a car accident, but low for an earthquake). A possible method to implement surprise is to look at the co-occurrence patterns of entity types. A headline which includes entities which do not usually occur together (e.g. *politician-celebrity*) might be more surprising that entity co-occurrences that are frequent (e.g. *politician-politician*). Methods which combine state-of-the-art NLP with Semantic Web resources could add depth and nuance to many research tasks. For example, an ongoing project on Active Video Watching (Dimitrova et al., 2017) uses NLP methods and ontologies to analyse user comments on videos.

Bibliography

Mohamed Ahmed, Stella Spagna, Felipe Huici, and Saverio Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 607–616, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1869-3.

Ioannis Arapakis, Berkant Barla Cambazoglu, and Mounia Lalmas. On the feasibility of predicting news popularity at cold start. In Luca Maria Aiello and Daniel McFarland, editors, *Social Informatics*, pages 290–299, Cham, 2014. Springer International Publishing.

Ioannis Arapakis, Filipa Peleja, Barla Berkant, and Joao Magalhaes. Linguistic benchmarks of online news article quality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1893–1902, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P16-1178`.

Ioannis Arapakis, Berkant Barla Cambazoglu, and Mounia Lalmas. On the feasibility of predicting popular news at cold start. *Journal of the Association for Information Science and Technology*, 68(5):1149–1164, 2017.

Yoav Artzi, Patrick Pantel, and Michael Gamon. Predicting responses to microblog posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 602–606, Montréal, Canada, June 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N12-1074`.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. Success with style: Using writing style to predict the success of novels. *Poetry*, 580(9):70, 2013.

Sitaram Asur and Bernardo a. Huberman. Predicting the Future with Social Media. In *2010*

IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pages 492–499, 2010.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

Roja Bandari, Sitaram Asur, and Bernardo Huberman. The pulse of news in social media: Forecasting popularity. In *International AAAI Conference on Web and Social Media*, 2012.

Chris Barr, editor. *The Yahoo! style guide: the ultimate sourcebook for writing, editing, and creating content for the digital world.* Macmillan, 2010.

Marco Toledo Bastos. Shares, pins, and tweets: News readership from daily papers to social media. *Journalism studies*, 16(3):305–325, 2015.

Monika Bednarek and Helen Caple. *News Discourse.* Continuum, 2012.

Monika Bednarek and Helen Caple. Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society*, 25(2):135–158, 2014.

Allan Bell. *The language of news media.* Blackwell Oxford, 1991.

Allan Bell. News time. *Time & Society*, 4(3):305–328, 1995.

Jonah Berger and Katherine L Milkman. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205, 2012.

Douglas Biber. *Variation across speech and writing.* Cambridge University Press, 1991.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python.* O'Reilly Media, Inc., 2009.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100, 2015.

Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *International AAAI Conference on Web and Social Media*, 2011.

Geert Brône and Seana Coulson. Processing deliberate ambiguity in newspaper headlines: Double grounding. *Discourse Processes*, 47(3):212–236, 2010.

Axel Bruns and Jean Burgess. Researching news discussion on twitter: New methodologies. *Journalism Studies*, 13(5-6):801–814, 2012.

Helen Caple and Monika Bednarek. Delving into the discourse: Approaches to news values in journalism studies and beyond. *Reuters Institute for the Study of Journalism*, 2013.

Anabela Carvalho. Media (ted) discourse and society: Rethinking the framework of critical discourse analysis. *Journalism Studies*, 9(2):161–177, 2008.

Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '14, pages 211–223, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2540-0.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 56–61, 2017.

Jan Chovanec. *Pragmatics of Tense and Time in News: From canonical headlines to online news texts*. Pragmatics & Beyond New Series. John Benjamins Publishing Company, 2014.

Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*, 1997.

Ryan Compton, David Jurgens, and David Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE, 2014.

David Conley and Stephen Lamble. *The Daily Miracle: An Introduction to Journalism*. Oxford University Press, 2006.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 249–260. ACM, 2013.

Colleen Cotter. *News talk: Investigating the language of journalism*. Cambridge University Press, 2010.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 892–901. Association for Computational Linguistics, 2012.

Maria Pia di Buono, Jan Šnajder, Bojana Dalbelo Basic, Goran Glavaš, Martin Tutek, and Natasa Milic-Frayling. Predicting news values from headline text and emotions. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 1–6, 2017.

Vania Dimitrova, Antonija Mitrovic, Alicja Piotrkowicz, Lydia Lau, and Amali Weerasinghe. Using learning analytics to devise interactive personalised nudges for active video watching. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 22–31. ACM, 2017.

Daniel Dor. On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 35 (5):695–721, 2003.

Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. Carmen: A twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, 2013.

Maeve Duggan, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. Social media update 2014. 2015. [Available at `http://www.pewinternet.org/2015/01/09/social-media-update-2014/`].

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P13-1174`.

Flavio Figueiredo, Jussara M Almeida, Fabrício Benevenuto, and Krishna P Gummadi. Does content determine information popularity in social media?: A case study of youtube videos' content and their popularity. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pages 979–982. ACM, 2014.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378, 1971.

Roger Fowler. *Language in the News*. Routledge, 1991.

Elfriede Fürsich. In defense of textual analysis: Restoring a challenged method for journalism and media studies. *Journalism Studies*, 10(2):238–252, 2009.

Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. Social Clicks: What and Who Gets Read on Twitter? In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, pages 179–192. ACM, June 2016.

Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34:443–498, 2009.

Johan Galtung and Mari Holmboe Ruge. The structure of foreign news the presentation of the Congo, Cuba and Cyprus Crises in four Norwegian newspapers. *Journal of Peace Research*, 2(1):64–90, 1965.

Herbert J Gans. *Deciding what's news: A study of CBS evening news, NBC nightly news, Newsweek, and Time*. Northwestern University Press, 1979.

Lorenzo Gatti, Gözde Özdal, Marco Guerini, Oliviero Stock, and Carlo Strapparava. Automatic creation of flexible catchy headlines. In *Proceedings of the Natural Language Processing meets Journalism Workshop*. International Joint Conference on Artificial Intelligence, 2016.

Jeffrey Gottfried and Elisa Shearer. News use across social media platforms 2016. *Pew Research Center, Washington, D.C.*, 2016. [Available at `http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/`; accessed 13th April 2018].

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English Gigaword LDC2003T05. Web Download. *Philadelphia: Linguistic Data Consortium*, 2003.

Mark Graham, Scott A Hale, and Devin Gaffney. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578, 2014.

Marco Guerini, Alberto Pepe, and Bruno Lepri. Do linguistic style and readability of scientific abstracts affect their virality? In *International AAAI Conference on Web and Social Media*, 2012.

Tony Harcup and Deirdre O'Neill. What is news? Galtung and Ruge revisited. *Journalism Studies*, 2(2):261–280, 2001.

Tony Harcup and Deirdre O'Neill. What is news? News values revisited (again). *Journalism Studies*, pages 1–19, 2016.

Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. Share, like, recommend: Decoding the social media news consumer. *Journalism Studies*, 13(5-6):815–824, 2012.

Kenneth Holmqvist, Jana Holsanova, Maria Barthelson, and Daniel Lundqvist. Reading or scanning? A study of newspaper and net paper reading. *Mind*, 2(3):4, 2003.

Jana Holsanova, Henrik Rahm, and Kenneth Holmqvist. Entry points and reading paths on newspaper spreads: comparing a semiotic analysis with eye-tracking measurements. *Visual Communication*, 5(1):65–93, 2006.

Chu-Cheng Hsieh, Christopher Moghbel, Jianhong Fang, and Junghoo Cho. Experts vs the crowd: Examining popular news prediction perfomance on Twitter. In *Proceedings of the World Wide Web Conference*, 2013.

Salman Jamali and Huzefa Rangwala. Digging digg: Comment mining, popularity prediction, and social network analysis. In *Proceedings of the International Conference on Web Information Systems and Mining (WISM 2009)*, pages 32–38. IEEE, 2009.

Karen S. Johnson-Cartee. *News narratives and news framing: Constructing political reality*. Rowman & Littlefield Publishers, 2005.

David Jurgens, Tyler Finethy, James McCorriston, Yi Xu, and Derek Ruths. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *International AAAI Conference on Web and Social Media*, 2015.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Hans Mathias Kepplinger and Simone Christine Ehmig. Predicting news decisions. An empirical test of the two-component theory of news selection. *Communications*, 31(1): 25–43, 2006.

Minkyoung Kim, Lexing Xie, and Peter Christen. Event diffusion patterns in social media. In *International AAAI Conference on Web and Social Media*, 2012.

Minkyoung Kim, David Newth, and Peter Christen. Modeling Dynamics of Diffusion Across Heterogeneous Social Networks: News Diffusion in Social Media. *Entropy*, 15 (10):4215–4242, 2013.

Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July 2003. Association for Computational Linguistics.

Shoubin Kong, Qiaozhu Mei, Ling Feng, Fei Ye, and Zhe Zhao. Predicting bursts and popularity of hashtags in real-time. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 927–930. ACM, 2014.

Sawa Kourogi, Hiroyuki Fujishiro, Akisato Kimura, and Hitoshi Nishikawa. Identifying attractive news headlines for social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1859–1862. ACM, 2015.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu. A large-scale sentiment analysis for yahoo! answers. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pages 633–642. ACM, 2012.

Jeffrey Kuiken, Anne Schuth, Martijn Spitters, and Maarten Marx. Effective headlines of newspaper articles in a digital environment. *Digital Journalism*, 5(10):1300–1314, 2017.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, 2010.

Himabindu Lakkaraju and Jitendra Ajmera. Attention prediction on social media brand pages. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 2157–2160. ACM, 2011.

Himabindu Lakkaraju, Julian J McAuley, and Jure Leskovec. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In *International AAAI Conference on Web and Social Media*, 2013.

Sara Leckner. Presentation factors affecting reading behaviour in readers of newspaper media: an eye-tracking perspective. *Visual Communication*, 11(2):163–184, 2012.

Kristina Lerman and Rumi Ghosh. Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *International AAAI Conference on Web and Social Media*, pages 90–97, 2010.

Kristina Lerman and Tad Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th International Conference on World Wide Web*, pages 621–630. ACM, 2010.

Ziming Liu. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of documentation*, 61(6):700–712, 2005.

Christoph Lofi, Joachim Selke, and Wolf-Tilo Balke. Information extraction meets crowdsourcing: a promising couple. *Datenbank-Spektrum*, 12(2):109–120, 2012.

Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the Association for Information Science and Technology*, 59(7):1041–1052, 2008.

Annie Louis and Ani Nenkova. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352, 2013.

Diana Maynard and Kalina Bontcheva. Challenges of evaluating sentiment analysis tools on social media. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), 2016.

Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*, pages 233–242. ACM, 2007.

Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Elisa Shearer. The Modern News Consumer: News attitudes and practices in the digital era. *Pew Research Center, Washington, D.C.*, 2016. [Available at `http://www.journalism.org/2016/07/07/the-modern-news-consumer/`; accessed 13th April 2018].

Kenny Olmstead, Amy Mitchell, and Tom Rosenstiel. Navigating news online: where people go, how they get there and what lures them away. *Pew Research Center, Washington, D.C.*, 2011. [Available at `http://www.journalism.org/2011/05/09/navigating-news-online/`; accessed 13th April 2018].

Deirdre O'Neill and Tony Harcup. News values and selectivity. *The Handbook of Journalism Studies*, pages 161–174, 2009.

Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

Heiko Paulheim and Johannes Fümkranz. Unsupervised generation of data mining features from linked open data. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, page 31. ACM, 2012.

Charles A Perfetti, Sylvia Beverly, Laura Bell, Kimberly Rodgers, and Robert Faux. Comprehending newspaper headlines. *Journal of Memory and Language*, 26(6):692–713, 1987.

Andrew Perrin. Social Media Usage 2005-2015. *Pew Research Center, Washington, D.C.*, 2015. [Available at `http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/`, accessed 13th April 2018].

Sasa Petrovic, Miles Osborne, and Victor Lavrenko. RT to Win! Predicting Message Propagation in Twitter. In *International AAAI Conference on Web and Social Media*, pages 586–589, 2011.

Angela Phillips. Sociability, speed and quality in the changing news environment. *Journalism Practice*, 6(5-6):669–679, 2012.

Greg Philo. Can discourse analysis successfully explain the content of media and journalistic practice? *Journalism Studies*, 8(2):175–196, 2007.

Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii, 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D08-1020`.

Amanda Potts, Monika Bednarek, and Helen Caple. How can computer-based methods help researchers to investigate news values in large datasets? *Discourse & Communication*, 9 (2):149–172, 2015.

Emily Pronin, Thomas Gilovich, and Lee Ross. Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. *Psychological Review*, 111(3):781, 2004.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A Comprehensive Grammar of the English Language*. Longman, 1985.

Dragomir Radev, Jahna Otterbacher, Adam Winkel, and Sasha Blair-Goldensohn. Newsinessence: summarizing online news topics. *Communications of the ACM*, 48(10):95–98, 2005.

Dragomir R Radev, Eduard Hovy, and Kathleen McKeown. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408, 2002.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria, 2013. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P13-1162`.

Julio Reis, Fabrício Benevenuto, Pedro Olmo, Raquel Prates, Haewoon Kwak, and Jisun An. Breaking the news: First impressions matter on online news. In *International AAAI Conference on Web and Social Media*, 2015.

Petar Ristoski and Heiko Paulheim. Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, 36:1–22, 2016.

Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. Where's@ wally?: a classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20. ACM, 2013.

Matthew Rowe and Harith Alani. Mining and comparing engagement dynamics across multiple social media platforms. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 229–238. ACM, 2014.

Parnia Samimi, Sri Devi Ravana, William Webber, and Yun Sing Koh. Effects of objective and subjective competence on the reliability of crowdsourced relevance judgments. *Information Research*, 22(1), 2017.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, 2008. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D08-1027`.

Le Hoang Son. Dealing with the new user cold-start problem in recommender systems: A comparative review. *Information Systems*, 58:87–104, 2016.

Kate Starbird and L Palen. (How) will the revolution be retweeted?: Information diffusion and the 2011 Egyptian uprising. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 7–16, 2012.

Justin Stewart, Homer Strong, Jeffery Parker, and Mark a. Bedau. Twitter Keyword Volume, Current Spending, and Weekday Spending Norms Predict Consumer Spending. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 747–753, 2012.

Gabor Szabo and Bernardo a. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80, 2010.

Terrence Szymanski, Claudia Orellana-Rodriguez, and Mark T. Keane. Helping news editors write better headlines: A recommender to improve the keyword contents and shareability of news headlines. In *Natural Language Processing meets Journalism Workshop*. International Joint Conference on Artificial Intelligence, 2016.

Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. In *ACL*, 2014.

Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias De Amorim, and Serge Fdida. From popularity prediction to ranking online news. *Social Network Analysis and Mining*, 4(1): 1–12, 2014.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

Giang Tran, Mohammad Alrifai, and Eelco Herder. Timeline summarization from relevant headlines. In *European Conference on Information Retrieval*, pages 245–256. Springer, 2015.

Manos Tsagkias, Wouter Weerkamp, and Maarten De Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1765–1768. ACM, 2009.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Web and Social Media*, pages 178–185, 2010.

Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 131–142. ACM, 2004.

Melissa Wall. Citizen journalism: A retrospective on what we know, an agenda for what we don't. *Digital Journalism*, 3(6):797–813, 2015.

James G Webster. *The marketplace of attention: How audiences take shape in a digital age*. MIT Press, 2014.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6): 80–83, 1945.

Sheng Yu and Subhash Kak. A survey of prediction using social media. *arXiv preprint arXiv:1203.1647*, pages 1–20, 2012.

Lun Zhang, Tai-Quan Peng, Ya-Peng Zhang, Xiao-Hong Wang, and Jonathan J.H. Zhu. Content or context: Which matters more in information processing on microblogging sites. *Computers in Human Behavior*, 31:242–249, 2014.

Arkaitz Zubiaga, Alex Voss, Rob Procter, Maria Liakata, Bo Wang, and Adam Tsakalidis. Towards real-time, country-level location classification of worldwide tweets. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):2053–2066, 2017.

Appendix A

Datasets

In the course of this thesis we produced the following datasets, which have been made publicly available:

- Piotrkowicz, A. (2017) Headlines corpora with automatically extracted news values scores. University of Leeds. [Dataset] https://doi.org/10.5518/147.

  This dataset includes two headlines corpora: *The Guardian* and *New York Times*. Each corpus consists of a unique headline identifier (to enable recreating the corpus by querying the relevant API) and news values scores for each headline.

- Piotrkowicz, A. (2017) Headlines data for social media popularity prediction. University of Leeds. [Dataset] https://doi.org/10.5518/174.

  This dataset includes all files used to build the global popularity prediction models in Chapter 8. The corpora include a unique headline identifier (to enable recreating the corpus by querying the relevant API), the extracted features (news values, linguistic style, metadata), and the corresponding news article popularity on Twitter and Facebook.

Appendix B

Examples of Preprocessed Headlines

Examples from *The Guardian*

Headline B.1. “Emma Watson’s makeup tweets highlight the commodification of beauty”

$H =$ { *Emma*, *Watson*, *’s*, *makeup*, *tweets*, *highlight*, *the*, *commodification*, *of*, *beauty* }

$C =$ { *makeup*, *tweets*, *highlight*, *commodification*, *beauty* }

$E =$ { EMMA WATSON, COMMODIFICATION }

Headline B.2. “Is the French prime minister Manuel Valls the new Tony Blair?”

$H =$ { *Is*, *the*, *French*, *prime*, *minister*, *Manuel*, *Valls*, *the*, *new*, *Tony*, *Blair*, *?* }

$C =$ { *Is*, *French*, *prime*, *minister*, *Manuel*, *Valls*, *new*, *Tony*, *Blair* }

$E =$ { FRENCH LANGUAGE, PRIME MINISTER OF FRANCE, PRIME MINISTER OF THE UNITED KINGDOM, MANUEL VALLS, TONY BLAIR }

Headline B.3. “Market feast”

$H =$ { *Market*, *feast* }

$C =$ { *Market*, *feast*}

$E =$ { }

Examples from *New York Times*

Headline B.4. "Piers Morgan Will Write for The Daily Mail"

$H =$ { *Piers*, *Morgan*, *Will*, *Write*, *for*, *The*, *Daily*, *Mail* }

$C =$ { *Piers*, *Morgan*, *Write*, *Daily*, *Mail* }

$E =$ { PIERS MORGAN, DAILY MAIL }

Headline B.5. "2014 Paris Motor Show: Audi Scores With the TT Sportback Concept"

$H =$ { *2014*, *Paris*, *Motor*, *Show*, *:*, *Audi*, *Scores*, *With*, *the*, *TT*, *Sportback*, *Concept* }

$C =$ { *Paris*, *Motor*, *Show*, *Audi*, *Scores*, *TT*, *Sportback*, *Concept* }

$E =$ { PARIS MOTOR SHOW, AUDI, AUDI TT, AUDI SPORTBACK CONCEPT, CONCEPT CAR }

Headline B.6. "'Fire!'"

$H =$ { *Fire*, *!* }

$C =$ { *Fire* }

$E =$ { }

Appendix C

Examples of Headlines Annotated with News Values

Table C.1: Examples of *The Guardian* headlines annotated with news values. Column headings refer to notation in Table 4.1

Headline	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**	**N14**	**N15**	**N16**
José Mourinho won't give Liverpool a free pass whatever his priorities	2	9397	12477	33	0	44	-2.13	0.88	0.25	0.25	0	0	0	1	26.3	0
Horse racing tips: Thursday 17 April	1	1031	891	7	0	18	-2	0	0	0	0	0	0	0	4.15	0.88
Will Chelsea Clinton run for the White House one day?	2	4549	11071	2	3.7	11	-2	0	0.14	0.57	0	0	0	1	4.15	0
Where are the UK's windfarms?	0	0	0	0	0	0	-1.88	0.38	0	0	0	0	0	0	2.14	0
Barry O'Farrell's resignation: the dos and don'ts of post-Icac etiquette	1	73	191	3	0	28	-1.75	1	0.5	0	0	0	0	1	0.36	0
Swansea City v Chelsea – as it happened, Nick Miller	3	11035	9875	23	0	69	-2	0	0.29	0.14	0	0	0	1	4.15	0.76

table continues

continue table

Headline	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**	**N14**	**N15**	**N16**
Martin Kaymer the Masters survivor driven by high ambition of Ryder Cup	2	475	1183	0	2.09	45	-1.75	0.75	0.44	0.11	0	0.08	0	0	0.82	0
War Horse's Jeremy Irvine to star in gay rights film Stonewall	4	7725	8999	2	0	49	-1.5	0.5	0.67	0.44	0	0	0	1	0.01	0
My guilty pleasure: Hitch	1	7	1	4	0	22	-2.5	0.75	0.67	0	0	0	0	0	799.54	0
The daily quiz, 16 April 2014	1	3	2	5	0	30	-2	0	0	0	0	0	0	1	43.09	0.73
South Korea ferry disaster: footage shows crew being rescued	1	7284	12108	7	0	9	-1.88	0.88	0.44	0.11	0	0	0	0	694.45	0
Getting creative work done with Oliver Burkeman	1	14	30	0	3.19	20	-1.63	0.88	0.67	0.17	0	0	0	1	0.92	0

table continues

continue table

Headline	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**	**N14**	**N15**	**N16**
Has Clive Palmer spent enough to win balance of power in the Senate?	4	550	637	7	0	38	-1.88	0.13	0.33	0.11	0	0	0.08	0	22.08	0
Teenage plane stowaway snuck aboard despite being caught on camera	1	182	14259	1	41.7	8	-1.88	0.63	0.13	0.13	0	0	0	0	0.01	0.57
Ben Watt - Hendra: exclusive album stream	1	113	168	0	0	20	-2.75	0.75	0.33	0	0	0	0	1	0.48	0
Gus Poyet hopes Sunderland find their form against Tottenham Hotspur	3	6931	4774	21	0	61	-1.75	0.5	0.25	0.25	0	0	0	1	6.04	0
Manchester City recover poise before final straight in title chase	1	5331	4487	29	0	18	-2.38	0.38	0.5	0.25	0	0	0	1	4.15	0

table continues

continue table

Headline	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**	**N14**	**N15**	**N16**
The Yashika Bageerathi case reveals the death of compassionate Conservatism	2	72	104	6	0	46	-1.75	0.25	0.29	0.29	0	0	0	0	1.58	0
Mike Ashley of Sports Direct buys 11% of House of Fraser	4	970	1632	10	10.67	91	-2	0	0.25	0.13	0	0	0	1	0.02	0.51
David Attenborough: changing viewing habits may halt future landmark series	1	2402	2199	0	0	12	-2.13	0.13	0.11	0.11	0	0	0	1	0	0

Table C.2: Examples of *New York Times* headlines annotated with news values. Column headings refer to notation in Table 4.1

Headline	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**	**N14**	**N15**	**N16**
Mexico: Band Member Is Found Dead	1	9161	9046	3	0	16	-2	0	0.17	0.17	0	0	0	1	4.04	0
Coalition Seeks to Send North Korea to International Court Over Rights Abuses	1	6697	4019	5	0	20	-2	0	0.1	0.3	0	0	0	0	4.04	0
Maybe Mother Isn't Losing It After All	0	0	0	0	0	0	-2.25	0.75	0.33	0.33	0	0.13	0	0	4.04	0
Murano, Italy, Still Sparkling After 700 Years	2	9515	6612	0	0	20	-2.13	0.13	0.25	0	0	0	0	0	4.04	0
Reports Tell of Scramble in Southwest Airlines Cockpit Before La Guardia Crash	1	1647	1624	0	0	15	-2	0	0.22	0.11	0	0	0	1	4.04	0

table continues

continue table

Headline	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16
Professors' Research Project Stirs Political Outrage in Montana	1	165	164	0	0	9	-2	0	0.14	0.29	0	0	0	1	4.04	0
Faces of Breast Cancer	1	4633	5332	0	0	9	-2.25	0.5	0	0	0	0	0	0	4.04	0
$16 Million Unit, Reserved by a Sponsor	0	0	0	0	0	0	-2	0	0.5	0	0	0	0	0	4.04	0.14
As Egyptians Grasp for Stability, Sisi Fortifies His Presidency	0	0	0	0	0	0	-2	0	0.67	0	0	0	0	0	0.06	0
Brooklyn District Attorney Will Ask Judge to Throw Out Murder Convictions	3	4561	4785	4	0	56	-2.5	0.75	0.33	0.11	0	0.09	0	1	0.01	0.24
Oklahoma Man Is Charged in Beheading of Co-Worker	1	2258	2409	1	0	18	-2	0.5	0.33	0.17	0	0	0	1	4.04	0

table continues

continue table

Headline	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**	**N14**	**N15**	**N16**
Latest Alaska Polls Show Surprising Shift Toward Mark Begich	2	5603	5570	0	4.17	41	-1.88	0.13	0.13	0.13	0.13	0	0	1	0.88	0
Obama Sees an Iran Deal That Could Avoid Congress	1	7011	5817	3	0	18	-1.25	0.75	0.17	0.33	0	0	0	0	4.04	0
Gerard Parkes, Actor on ‘Fraggle Rock,’ Dies at 90	2	554	797	2	0	14	-2	0	0.5	0	0	0	0	1	4.04	0
Bayern Munich Cruises Past Roma in Champions League	3	13793	26993	0	0	68	-2	0	0.29	0	0	0	0	0	4.04	0
Tesco Chairman to Step Down as Overstatement of Profit Grows	1	2198	6019	0	5.38	13	-2	0.25	0.33	0	0	0	0	0	0.09	0
Wall St. Extends Rally on Solid Corporate Results	0	0	0	0	0	0	-1.63	1.38	0.29	0	0	0	0	0	10.64	0

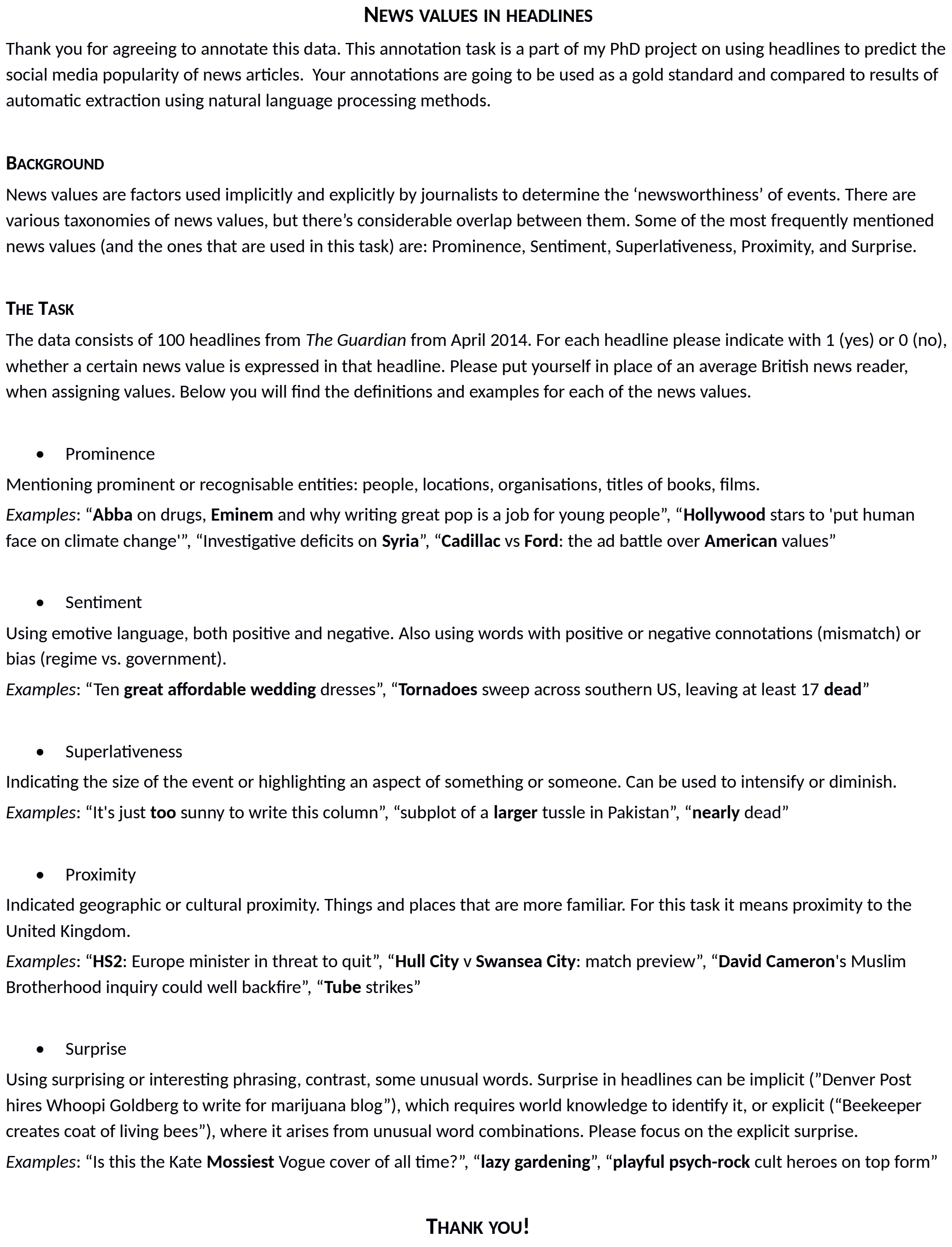
table continues

continue table

Headline	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**N9**	**N10**	**N11**	**N12**	**N13**	**N14**	**N15**	**N16**
Restoring a One-of-a-Kind Corvette Collection	2	1097	1301	0	0	36	-2	0	0.5	0	0	0	0	0	4.04	0
The Cost of Campaigns	0	0	0	0	0	0	-2	0	0.5	0.5	0	0	0	0	4.04	0
An Opera Under Fire	0	0	0	0	0	0	-2	0	0	0	0	0	0	0	4.04	0

Appendix D

Instructions for Obtaining News Values Gold Standard



NEWS VALUES IN HEADLINES

Thank you for agreeing to annotate this data. This annotation task is a part of my PhD project on using headlines to predict the social media popularity of news articles. Your annotations are going to be used as a gold standard and compared to results of automatic extraction using natural language processing methods.

BACKGROUND

News values are factors used implicitly and explicitly by journalists to determine the ‘newsworthiness’ of events. There are various taxonomies of news values, but there’s considerable overlap between them. Some of the most frequently mentioned news values (and the ones that are used in this task) are: Prominence, Sentiment, Superlativeness, Proximity, and Surprise.

THE TASK

The data consists of 100 headlines from *The Guardian* from April 2014. For each headline please indicate with 1 (yes) or 0 (no), whether a certain news value is expressed in that headline. Please put yourself in place of an average British news reader, when assigning values. Below you will find the definitions and examples for each of the news values.

- Prominence

Mentioning prominent or recognisable entities: people, locations, organisations, titles of books, films.

Examples: “**Abba** on drugs, **Eminem** and why writing great pop is a job for young people”, “**Hollywood** stars to 'put human face on climate change'”, “Investigative deficits on **Syria**”, “**Cadillac** vs **Ford**: the ad battle over **American** values”

- Sentiment

Using emotive language, both positive and negative. Also using words with positive or negative connotations (mismatch) or bias (regime vs. government).

Examples: “Ten **great affordable wedding** dresses”, “**Tornadoes** sweep across southern US, leaving at least 17 **dead**”

- Superlativeness

Indicating the size of the event or highlighting an aspect of something or someone. Can be used to intensify or diminish.

Examples: “It's just **too** sunny to write this column”, “subplot of a **larger** tussle in Pakistan”, “**nearly** dead”

- Proximity

Indicated geographic or cultural proximity. Things and places that are more familiar. For this task it means proximity to the United Kingdom.

Examples: “**HS2**: Europe minister in threat to quit”, “**Hull City** v **Swansea City**: match preview”, “**David Cameron**'s Muslim Brotherhood inquiry could well backfire”, “**Tube** strikes”

- Surprise

Using surprising or interesting phrasing, contrast, some unusual words. Surprise in headlines can be implicit (”Denver Post hires Whoopi Goldberg to write for marijuana blog”), which requires world knowledge to identify it, or explicit (“Beekeeper creates coat of living bees”), where it arises from unusual word combinations. Please focus on the explicit surprise.

Examples: “Is this the Kate **Mossiest** Vogue cover of all time?”, “**lazy gardening**”, “**playful psych-rock** cult heroes on top form”

THANK YOU!

Appendix E

Survey Used on Crowdflower

Hello and thanks for participating in this task!

Instructions

First we'll ask you to answer a few preliminary questions about you.

The task is then split into **two parts.**

PART 1: Headline popularity
- You will see **48 headlines** from The Guardian (a major British newspaper).
- For each headline, please **indicate *how likely is it that other people will click on this headline***. This means judging whether a given headline will have a wide audience appeal.
-

PART 2: Judgement criteria
- You will see **12 features of headlines**.
- Please **indicate to what extent each feature influences your decision** about clicking on headlines.

You will have a chance to provide some feedback about this task. We really appreciate your comments.

At the **end of the survey** you will be given a **code** that you need to **enter back on the CrowdFlower** website for this task in order to mark it as complete. Thank you very much and enjoy!

Please enter your CrowdFlower contributor ID:

PRELIMINARIES

Q1 What is your age?
- Under 18
- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65 - 74
- 75 - 84
- 85 or older

Q2 What is your gender?
- Male
- Female
- Other

Q3 In which country do you currently reside?
[single choice from a list of 195 countries]

Q4 Is English your native language?
- Yes
- No

Answer If Is English your native language? No Is Selected
Q4.1 What is your native language?
[single choice from a list of 72 languages]

Q5 How often do you read news online?
- Daily
- 4-6 times a week
- 2-3 times a week
- Once a week
- Never

PART 1: HEADLINES (*The Guardian* version)

Q6.1 How likely is it that other people will click on this headline?

Question	Extremely likely	Slightly likely	Neutral	Slightly unlikely	Extremely unlikely
Reviews roundup: your top reads this month					
Twitter buys UK 'social TV' firm SecondSync					
Chinese state energy firm ups shale gas spend to £950m					
Care homes are not all dreary TV dungeons that smell of wee					
In praise of butter					
Marc Platt obituary					
How can charities balance innovation and risk? - live discussion					
Scottish bird of prey colony hit by mass poisonings					
A puddle to which hundreds of bright birds poured down like blossom					
Top 10 daftest ways to become a world champion					
Atari's ET: which video games deserve to be buried in the desert?					
The Co-op is an idea worth fighting for					

Q6.2 How likely is it that other people will click on this headline?

	Extremely likely	Slightly likely	Neutral	Slightly unlikely	Extremely unlikely
Readers recommend: eccentric songs - results					
Lexi Thompson wins maiden major crown after Kraft Nabisco triumph					
Melbourne man accused of murdering partner 'had family violence order'					
Palestinian statehood bid may derail Middle East peace process					
John Kerry discusses Ukraine crisis with Russian foreign minister					
A Rational Fear: Who is Mike Baird? - video					
Kigali's future or costly fantasy? Plan to reshape Rwandan city divides opinion					
Oscar Pistorius trial: spokeswoman denies athlete took acting lessons					
Restricting onshore windfarms would be a costly policy decision					
Does Westminster have a problem with women?					
Global Youth Index shows young people worldwide have a rough deal					
Roberto Martinez: 'David Moyes will bounce back from sacking'					

Q6.3 How likely is it that other people will click on this headline?

	Extremely likely	Slightly likely	Neutral	Slightly unlikely	Extremely unlikely
Children 'kept from parents' at centre for failed asylum seekers					
Purple haze: Coachella festival seen through an infrared lens - in pictures					
How to make a wildlife-friendly garden					
Equal pay awakenings: when did you realise you were underpaid?					
Chicken schnitzel with herbs and parmesan - Bondi Harvest video recipe					
Teachers to vote on strike motion					

Bodies of Light by Sarah Moss review - 'a hard-working novel about hard-working women'					
No-fly list used by FBI to coerce Muslims into informing, lawsuit claims					
Saracens' Owen Farrell finally finds range to thwart 14-man Ulster					
Scottish independence would damage maritime defence, says First Sea Lord					
Epigenetics 101: a beginner's guide to explaining everything					
China's 'eco-cities': empty of hospitals, shopping centres and people					

Q6.4 How likely is it that other people will click on this headline?

	Extremely likely	Slightly likely	Neutral	Slightly unlikely	Extremely unlikely
#AustraliansForCoal is the latest sign of an industry in values freefall					
Wigan v Arsenal: FA Cup semi-final - as it happened					
Can Cortana and other new features turn Windows Phone around?					
Scottish roundup: Richard Brittain's goal puts Kilmarnock in the mire					
Journey to North Korea's volcano: British scientists visit Mount Paektu					
Ask a grown-up: why do old people have grey hair?					
Restaurant Wars: The Battle for Manchester; The Wonder of Bees - TV review					
Stinkbomb & Ketchup-Face and the Badness of Badgers by John Dougherty - review					
Ideas for 16-17 April					
The Joy of Six: David Boon					
Five ways to tone up your torso					
Prince George has a play date in New Zealand - in pictures					

PART 1: HEADLINES (*New York Times* version)

Q6.1 How likely is it that other people will click on this headline?

	Extremely likely	Slightly likely	Neutral	Slightly unlikely	Extremely unlikely
Museums Plug In					
The Big Sheet on the Wall Enters the 21st Century					
Two Danish Badminton Players Report a Fixing Invitation					
An Apple a Day, for 47 Years					
Policy Change Could Benefit New York's Landlords and Tenants					
Weekend Auto Calendar: Vintage Racing Meets Southern Charm					
Britain Pledges Millions to Fight Ebola and Chides Others to Spend More					
Deal Saves Historic Nashville Studio					
Is That Cut a Tri-Tip or What?					
From a Rwandan Dump to the Halls of Harvard					
Rachel Hock, Paul Mysliwiec					
Cooking With Cauliflower, a Feisty Vegetable That					

Can Take a Punch

Q6.2 How likely is it that other people will click on this headline?

	Extremely likely	Slightly likely	Neutral	Slightly unlikely	Extremely unlikely
Environment Is Grabbing Big Role in Ads for Campaigns					
New Freedoms in Tunisia Drive Support for ISIS					
The Rebirth of Tijuana					
In New York, Protections Offered for Medical Workers Joining Ebola Fight					
Data-Driven Campaigns Zero In on Voters, but Messages Are Lacking					
Harrell, South Carolina House Speaker, Pleads Guilty					
Autumn Is in the Air, but for Marketers, Christmas Has Already Begun					
The Making of Saints, in Africa and Beyond					
F.A.A. Tells Airlines to Replace Some Boeing Cockpit Displays					
Anna Selberg and Colin Samuels					
Web-Era Trade Schools, Feeding a Need for Code					
Ballot Item Would Reform Redistricting, at Least in Theory					

Q6.3 How likely is it that other people will click on this headline?

	Extremely likely	Slightly likely	Neutral	Slightly unlikely	Extremely unlikely
The Collapse of the Secret Service					
Lawmakers Grill French Candidate for European Economic Post					
Ebola Patient Sent Home Despite Fever, Records Show					
Lights, Catcher, Action!					
Genocide Trial Begins for Khmer Rouge Leaders					
Roasted Pepper Tartine					
Oklahoma Man Is Charged in Beheading of Co-Worker					
Kris Kobach Pushed Kansas to the Right. Now Kansas Is Pushing Back.					
The Cost of Campaigns					
The Meaning of Fulfillment					
Smell Turns Up in Unexpected Places					
Jazz Listings for Oct. 3-9					

Q6.4 How likely is it that other people will click on this headline?

	Extremely likely	Slightly likely	Neutral	Slightly unlikely	Extremely unlikely
A Choreographer Drawn to Change					
San Quentin’s Giants					
A Rational Quarantine					
OPEC Split as Oil Prices Fall Sharply					

Social Security Benefits to Rise Slightly Again

In Shorter-Game Experiment, Nets See Disparities as Minute

Deportation Up in 2013; Border Sites Were Focus

By the Sea, You and Me

Freezing Your Eggs: When, If and Why

A Prescription for Life's Final Stretch

Combating a Flood of Early 401(k) Withdrawals

A Crisis, an Alias and a Cocktail With Quince

PART 2: JUDGEMENT CRITERIA (*The Guardian* version)

Q7.1 To what extent does the following feature influence you in your choice of headlines?

PROMINENCE: Mentioning prominent or recognisable entities like people, locations, organisations, titles, characters, etc.

Examples:
“*Abba* on drugs, *Eminem* and why writing great pop is a job for young people”
“*Hollywood* stars to 'put human face on climate change'”
“Investigative deficits on *Syria*”
“*Cadillac* vs *Ford*: the ad battle over *American* values”

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think prominence influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider prominence when clicking on headlines					

Q7.2 To what extent does the following feature influence you in your choice of headlines?

SENTIMENT: Using emotive language, both positive and negative. Also using words with positive or negative connotations (e.g. mismatch) or bias (e.g. regime vs. government).

Examples:
“Ten great affordable wedding dresses”
“*Tornadoes* sweep across southern US, leaving at least 17 *dead*”

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think sentiment influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider sentiment when clicking on headlines					

Q7.3 To what extent does the following feature influence you in your choice of headlines?

MAGNITUDE: Indicating the size of the event or highlighting an aspect of something or someone. Can be used to intensify or diminish.

Examples:
“It's just *too* sunny to write this column”
“subplot of a *larger* tussle in Pakistan”
“*nearly* dead”

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think magnitude influences OTHER PEOPLE to click on headlines					

I PERSONALLY consider magnitude when clicking on headlines					

Q7.4 To what extent does the following feature influence you in your choice of headlines?
PROXIMITY: Indicating geographic or cultural proximity. Things and places that are more familiar.

Examples (for a British reader):
"*HS2*: Europe minister in threat to quit"
"*Hull City* v *Swansea City*: match preview"
"*David Cameron*'s Muslim Brotherhood inquiry could well backfire"
"*Tube* strikes"

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think proximity influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider proximity when clicking on headlines					

Q7.5 To what extent does the following feature influence you in your choice of headlines?

SURPRISE: Using surprising or interesting phrasing, contrast, some unusual words.

Examples:
"Is this the Kate *Mossiest* Vogue cover of all time?"
"*lazy gardening*"
"*playful psych-rock* cult heroes on top form"

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think surprise influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider surprise when clicking on headlines					

Q7.6 To what extent does the following feature influence you in your choice of headlines?

BREVITY: Short, concise headlines.

Example: "Fashion really does matter"

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think brevity influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider brevity when clicking on headlines					

Q7.7 To what extent does the following feature influence you in your choice of headlines?

SIMPLICITY: Using simple grammar and vocabulary.

Example: "What makes the perfect burger?"

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think simplicity influences OTHER PEOPLE to click on headlines					

I PERSONALLY consider simplicity when clicking on headlines					

Q7.8 To what extent does the following feature influence you in your choice of headlines?

UNAMBIGUITY: Using grammar and vocabulary which points to just one meaning.

Example: "George Osborne announces export credit scheme"

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think unambiguity influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider unambiguity when clicking on headlines					

Q7.9 To what extent does the following feature influence you in your choice of headlines?

PUNCTUATION: Using punctuation like exclamation marks, question marks, quote marks.

Example: "Is the 'cost of living crisis' over?"

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think punctuation influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider punctuation when clicking on headlines					

Q7.10 To what extent does the following feature influence you in your choice of headlines?

NUMBER OF NOUNS: Using many nouns (e.g. tree, bank, president).

Example: "Shock as Co-op announces electricity price rise"

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think number of nouns influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider number of nouns when clicking on headlines					

Q7.11 To what extent does the following feature influence you in your choice of headlines?

NUMBER OF VERBS: Using many verbs (e.g. say, attack, die).

Examples: "Want a promotion? Then be seen and be heard"

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think number of verbs influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider number of verbs when clicking on headlines					

Q7.12 To what extent does the following feature influence you in your choice of headlines?

NUMBER OF ADVERBS: Using many adverbs (e.g. early, greatly, finally, now).

Examples:
"Booming Britain running *smoothly*?"
"More Australian teenagers are *just* saying no to alcohol"

	Definitely yes	Probably yes	Might or might not	Probably not	Definitely not
I think number of adverbs influences OTHER PEOPLE to click on headlines					
I PERSONALLY consider number of adverbs when clicking on headlines					

PART 2: JUDGEMENT CRITERIA (*New York Times* version)

Same as *The Guardian* version with the exception of:

Q7.4 To what extent does the following feature influence you in your choice of headlines?

PROXIMITY: Indicating geographic or cultural proximity. Things and places that are more familiar.

Examples (for an American reader):
"First Ebola Case Found in the *U.S.*"
"Unwelcome Visitors to the *White House*"
"*Orioles'* Powerful Lineup"

Thank you very much for completing this task! We hope you enjoyed and that it has given you some insights about how you select headlines.

The code you need to enter on CrowdFlower is: HEADLINES_ROCK

Please enter your CrowdFlower contributor ID:

If you have any comments or feedback about this task, please enter it here:

Appendix F

Comparison of State-of-the-Art Features for ‘Cold-Start’ News Article Popularity Prediction

Feature type	Bandari et al. (2012)	Notes	Arapakis et al. (2014)	Notes	Our features
Prominence	number of named entities	Entity recognition using Stanford NER (place, person, organisation).	number of entities	Entity recognition using in-house software.	number of entities
			Wikipedia popularity	Prominence of entities in the title and article body is calculated separately.	Wikipedia long-term prominence
					Wikipedia day before prominence
	maximum entity score	Scores calculates using historical prominence on Twitter over a month normalised by the number of articles.	Twitter popularity	Summed popularity on Twitter and Web search calculated one hour, one day, and one week before article's publication.	
	average entity score		Web search popularity		
					Wikipedia burstiness

table continues

continue table

Feature type	**Bandari et al. (2012)**	**Notes**	**Arapakis et al. (2014)**	**Notes**	**Our features**
					Wikipedia current burst size
					News source recent prominence
Sentiment	subjectivity	Binary feature. Used subjectivity classifier from Ling-Pipe.	sentimentality score	Used SentiStrength positive/negative scores for individual sentences.	sentiment
			polarity score		polarity
					connotations
					bias
Magnitude					comparative/ superlative
					intensifiers
					downtoners
Proximity					proximity
Surprise					surprise

table continues

continue table

Feature type	**Bandari et al. (2012)**	**Notes**	**Arapakis et al. (2014)**	**Notes**	**Our features**
Uniqueness					uniqueness
Brevity			number of words	Calculated for title and article body.	number of words
			number of characters		number of characters
			number of sentences	Calculated only for article body.	
Simplicity					parse tree height
					number of non-terminal tree nodes
					entropy
					information content
					word frequency
Unambiguity					modality
					number of senses

table continues

continue table

Feature type	Bandari et al. (2012)	Notes	Arapakis et al. (2014)	Notes	Our features
Punctuation					exclamation mark
					question mark
					quote marks
Nouns					headlinese
					proportion of noun phrases
			proportion of nouns		proportion of nouns
					proportion of proper nouns
Verbs			proportion of verbs		proportion of verbs
					proportion of verb phrases
Adverbs			proportion of adverbs		proportion of adverbs

Appendix G

SPARQL Queries for Relating TagMe Entities to Countries

Query to Obtain Wikidata ID for a TagMe Entity

```
PREFIX wd:  <http://www.wikidata.org/entity/>
SELECT ?s
WHERE {
    ?s ?p "United␣Kingdom"@en .
}
LIMIT 1
```

Listing G.1: SPARQL query for obtaining Wikidata IDs using the entity *United Kingdom* as an example

Queries to Connect Entities to US/UK

These queries check whether a given entity (represented by a Wikidata ID) connects to the US or UK node in the Wikidata graph in fewer than three edges.

One edge to UK/US node

```
PREFIX wd: <http://www.wikidata.org/entity/>
ASK
WHERE {
        WIKI_ID ?p wd:Q145 .
}
```

Listing G.2: SPARQL query for connection to UK with one edge

Two Edges to UK/US Node

```
PREFIX wd: <http://www.wikidata.org/entity/>
ASK
WHERE {
        WIKI_ID ?p ?o .
    ?o ?p wd:Q145 .
}
```

Listing G.3: SPARQL query for connection to UK with two edges (same property)

```
PREFIX wd: <http://www.wikidata.org/entity/>
ASK
WHERE {
        WIKI_ID ?p1 ?o .
    ?o ?p2 wd:Q145 .
}
```

Listing G.4: SPARQL query for connection to UK with two edges (two different properties)

Three Edges to UK/US Node

```
PREFIX wd: <http://www.wikidata.org/entity/>
ASK
WHERE {
        WIKI_ID ?p ?o1 .
        ?o1 ?p ?o2 .
    ?o2 ?p wd:Q145 .
}
```

Listing G.5: SPARQL query for connection to UK with three edges (same property)

```
PREFIX wd: <http://www.wikidata.org/entity/>
ASK
WHERE {
        WIKI_ID ?p1 ?o1 .
        ?o1 ?p2 ?o2 .
    ?o2 ?p3 wd:Q145 .
}
```

Listing G.6: SPARQL query for connection to UK with three edges (different properties)