Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

Headlines Matter: Using Headlines to Predict the Popularity of News Articles on Twitter and Facebook

Alicja Piotrkowicz
Vania Dimitrova
School of Computing
University of Leeds, UK

Jahna Otterbacher
Social Information Systems
Open University of Cyprus

Katja Markert
Institut für Computerlinguistik
Universität Heidelberg, Germany

Abstract

Social media like Facebook or Twitter have become an entry point to news for many readers. In that scenario, the headline is the most prominent – and often the only visible – part of the news article. We propose a novel task of using *only headlines* to predict the popularity of news articles. The prediction model is evaluated on headlines from two major broadsheet news outlets – *The Guardian* and *New York Times*. We significantly improve over several baselines, noting differences in the model performance between Facebook and Twitter.

Introduction and Related Work

Headlines are vital in both capturing readers' attention and in influencing their online reading experience of news. In fact, approximately six in 10 people limit their reading to headlines only, without clicking on a link to the full article[1]. Eye-tracking studies have confirmed this behaviour empirically; 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 (Holsanova, Rahm, and Holmqvist 2006). Furthermore, there are many online spaces where headlines are the only visible part of the news article; for example news feeds and social media.

Yet despite this, headlines have not been considered before as the sole source of data for news article popularity prediction. Most models make use of post-publication data, such as the number of early adopters (Castillo et al. 2014). These methods model *popularity development*, e.g. they might use the number of tweets within the first hour *after article publication* to predict later or final popularity.

On the other hand, approaches which tackle what Arapakis, Cambazoglu, and Lalmas (2014) call the "cold start problem", i.e. the prediction of news article popularity *prior to publication*, are still in their infancy. In particular, these approaches offer limited insight into which aspects of the news article text make it popular online. Bandari, Asur, and Huberman (2012) use a small number of text features related to topical category, named entities' prominence and sentiment. Arapakis, Cambazoglu, and Lalmas (2014) reproduce and improve upon the work by Bandari, Asur, and Huberman (2012). They also add a small number of linguistic and prominence features, but their main focus is on evaluation methods. Both Bandari, Asur, and Huberman (2012) and Arapakis, Cambazoglu, and Lalmas (2014) use the news source as a feature, which is shown to be the overwhelming determiner of popularity. However, if the newsroom staff want to adjust article content to reach larger audiences, this is unhelpful, as news source is out of their control. Moreover, these previous models consider headlines and article body jointly. As headlines play a crucial role in the online news domain, it is worth investigating to what extent we can predict an article's popularity from the headline alone. Our goal is to investigate a wide variety of text features extracted from headlines and determine whether they have impact on social media popularity of news articles. We enhance prior work by: (i) using only headlines; (ii) introducing new features; and (iii) using a source-internal evaluation.

Data Collection

We created two corpora of news headlines and obtained the social media popularity for each headline.

News corpora. We used two major broadsheet newspapers — *The Guardian* and *New York Times*. We downloaded all headlines published during April 2014 (Guardian training), July 2014 (Guardian test), October 2014 (NYT training), and December 2014 (NYT test)[2]. Table 1 includes some example headlines with their popularity scores.

Social media data. We measure a news article's social media popularity by the number of times it is cited on Twitter and Facebook. The article URL was used as the search query for the Twitter Search API[3] to obtain the number of tweets and retweets one, three, and seven days after the article's publication. The process was repeated for Facebook likes and shares using the Facebook FQL API.[4]

Popularity measures. Tweets and retweets, as well as shares and likes, are combined into two metrics: Twitter and Facebook popularity. We found that in our datasets Twitter and Facebook popularity after three and seven days did not differ significantly, and so throughout the paper we report

[1] http://bit.ly/21LwfS5

[2] *The Guardian* data: Guardian Content API, *New York Times*: NYT Article Search API.

[3] https://dev.twitter.com/docs/api/1.1/get/search/tweets

[4] https://developers.facebook.com/docs/technical-guides/fql/

Table 1: Examples of the most and least popular headlines.

	The Guardian	*New York Times*
Most popular	"Capitalism simply isn't working and here are the reasons why" (T=2299, F=23840)	"Doctor in New York City Is Sick With Ebola" (T=12780, F=46603)
Least popular	"Corrections and clarifications" (T=0, F=0) "Fears misplaced over letting Lords resign" (T=5, F=0)	"Pastis and Ouzo: The Soccer of Liquors" (T=0, F=0) "Alaska's Political Outlook" (T=0, F=1)

popularity after three days, yielding two social media popularity measures: T = Twitter popularity after three days, and F = Facebook popularity after three days.

Data overview. The popularity measures show a strongly Zipfian distribution. Twitter and Facebook measures correlate well with each other (Guardian: ρ=0.74, NYT: ρ=0.6). 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 (in 2016 Facebook had 1.7 billion active users to Twitter's 0.3 billion[5]). News source also plays an important role, as *New York Times* articles are more often shared on social media (this follows the finding by Bandari, Asur, and Huberman (2012) that news source is the strongest predictor of social media popularity of news articles).

Headline Features

We use two types of features: journalism-inspired *news values* and *linguistic style*. Details of feature implementations are outlined in Piotrkowicz, Dimitrova, and Markert (2017).

News Values

News values, a concept originating in journalism studies, refer to aspects of news stories which make them newsworthy. While there are many news values taxonomies, there is considerable overlap (cf. Caple and Bednarek (2013)) and we implement six news values which are frequently included.

Prominence. Reference to prominent entities is one of the key news values. We approximate Prominence as the amount of online attention an entity gets. We extend previous work by using wikification for obtaining entities, which ensures a wide variety of entity types. We implement six Prominence features: (i) number of wikified entities; (ii) news recent prominence (the number of entity mentions in headlines of the relevant news outlet); (iii) long-term prominence (the median number of daily Wikipedia page views over a year for an entity); (iv) day-before prominence (number of Wikipedia page views for an entity the day before a given headline was published); (v) current burst size (if an entity is 'bursty', how much above the average are the entity's page views); and (vi) burstiness (if an entity is 'bursty', how many times over a year the entity has been in a burst).

Sentiment. Sentiment refers to both negative and positive vocabulary. We calculate four features using SentiWordNet (Baccianella, Esuli, and Sebastiani 2010) scores: (i) sentiment ($maxPos - maxNeg - 2$); (ii) polarity ($maxPos + maxNeg$); (iii) proportion of biased words; and (iv) proportion of positively/negatively connotated words.

Magnitude. This refers to the size or impact of a news event. There are three features: (i) proportion of comparative and superlative words (based on POS tags); (ii) proportion of intensifiers; and (iii) proportion of downtoners.

Proximity. We focus on geographic proximity to the news source, which assumes that readers from the same country as the news outlet constitute a large part of its readership. We implement Proximity as an explicit reference to news outlet country (UK/US related keywords) in the headline text.

Surprise. Surprising headlines draw attention. We measure surprise by calculating the commonness of syntactic chunks in a headline with reference to a Wikipedia corpus[6].

Uniqueness. Headlines should be novel. To investigate this, for a given headline we go through recent past headlines to see if there are any highly similar ones. For a pair of headline and past headline vectors (created using a *tf-idf* weighted Gigaword corpus) we calculate cosine similarity. The highest cosine similarity is assigned as the feature value.

Style

For linguistic style we calculate features inspired by journalism studies and NLP work on the effect of phrasing.

Brevity. Headlines need to be short. We implement Brevity as the number of tokens and number of characters.

Simplicity. Easy-to-understand headlines use simple syntax and vocabulary. We use two syntactic complexity features: (i) parse tree height, and (ii) number of non-terminal tree nodes. To measure lexical complexity we implement four features: (i) entropy (calculated using a trigram language model built using the CMU-Cambridge Toolkit on the *New York Times* section of the Gigaword corpus); (ii) proportion of difficult words (any word not occurring among the 5000 most common words in the language model); (iii) median word frequency (using the unlemmatised word frequency lists[7]); and (iv) information content (calculated for nouns and verbs on British National Corpus).

Unambiguity. News text should not be ambiguous. We use two features for headline ambiguity: median number of senses per word from WordNet, and modality (modal event or modal relation between events; using TARSQI[8]).

Punctuation. *The Guardian*'s style guide for headlines[9]

[5]http://bit.ly/2ddRJHi

[6]http://www.nlp.cs.nyu.edu/wikipedia-data

[7]http://www.wordfrequency.info/; British National Corpus for Guardian, Corpus of Contemporary American English for NYT.

[8]http://www.timeml.org/site/tarsqi/toolkit/index.html

[9]http://www.theguardian.com/guardian-observer-style-guide-h

discourages using quote, question, and exclamation marks. We implement binary features indicating their presence.

Nouns. *The Guardian*'s style guide cautions against using too many successive nouns (so-called 'headlinese'). We implement four features: (i) three consecutive nouns (binary feature); (ii) number of noun phrases; (iii) proportion of common nouns; and (iv) proportion of proper nouns.

Verbs. Using verbs is encouraged in headlines in *The Guardian*'s style guide. We implement two features: (i) number of verb phrases, and (ii) proportion of verbs.

Adverbs. Adverbs, especially adverbs of manner, are frequently used in headlines. We use the proportion of adverbs.

Prediction Model

Using these features, our goal is to predict the popularity of news articles on Twitter and Facebook from the article's headline. We build separate prediction models for each news source, thus avoiding news source popularity effects.

Method

We used regression (support vector regression with RBF kernel). Arapakis, Cambazoglu, and Lalmas (2014) argue that using classification for popularity prediction is not appropriate, as class splits potentially introduce bias towards articles with low popularity. Popularity measures – T, F – were log-transformed in order to improve model fit.

Results were evaluated on the test set[10]. Two evaluation metrics were used: Kendall's tau rank correlation coefficient (τ) and mean absolute error (MAE). Significance testing was performed using z-test for τ and t-test for MAE.

Baselines

We used three baselines: a unigrams baseline and two state-of-the-art baselines. Our model's features are denoted as $\mathcal{M}$.

Unigrams ($\mathcal{M}_\mathcal{U}$). We used 1000 most frequent unigrams.

State-of-the-art reimplementations: Bandari, Asur, and Huberman (2012) ($\mathcal{M}_\mathcal{B}$) and Arapakis, Cambazoglu, and Lalmas (2014) ($\mathcal{M}_\mathcal{A}$) originally used full article text, but we ran these baselines on the same dataset as ours (i.e. headlines only). We aimed at as close a reimplementation as possible, but in some cases we had to make adjustments. We used Stanford Named Entity Recognizer and SentiWordNet for Prominence and Sentiment features, respectively. Without access to archival Twitter data, we used Wikipedia to calculate Prominence features. Finally, unlike in the original implementations there is no news source feature (as our goal is a source-internal evaluation).

The two state-of-the-art approaches, as well as similar tasks (Lakkaraju, McAuley, and Leskovec 2013), make use of metadata that is available at the time of article publication (category, time). The reimplemented baselines and our full model ($\mathcal{M}$) also include metadata. Following the implementation by Arapakis, Cambazoglu, and Lalmas (2014), both category and publication date and time are implemented as binary features in our model. Bandari, Asur, and Huberman (2012) calculate a category score ($\frac{\#citations\ per\ category}{\#articles\ per\ category}$).

[10]Evaluation using cross-validation is not appropriate, because the data is temporally ordered and one of our features, headline uniqueness, makes use of the temporal ordering.

Results and Discussion

We report regression results against different baselines.

Table 2: Regression results of baselines against our model ($\mathcal{M}$) using all features (news values, style, metadata). Result in bold indicates improvement of $p<0.05$.

	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**

Table 3: Regression results of baselines against our model ($\mathcal{M}$) using headline features only (news values and style). Result in bold indicates improvement of $p<0.05$.

	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**

Performance against baselines using full feature set (Table 2). Our model ($\mathcal{M}$) significantly outperforms the baselines for nearly all measures. The exceptions are the MAE results for Twitter in the NYT dataset, where unigrams baseline outperforms the model. However, for that same dataset our model achieves significantly higher correlations. The best results were achieved for Twitter in the Guardian dataset (τ=0.43, MAE=0.68). This is a promising result, considering that this is the first attempt to use headlines for a news article popularity prediction task.

Performance against baselines using content features only (Table 3). When limited to features that can be extracted directly from headline text (news values and style), our model shows a considerably better improvement for most measures than when comparing models with metadata (improvement in correlation of approx. 40%, compared to 5-10% when using metadata). The highest correlation was achieved for Twitter in the Guardian dataset ($\tau = 0.29$), and the lowest MAE for Twitter in NYT dataset (MAE=0.69).

Performance of feature groups (Table 4). Using all features significantly outperforms any individual feature group at $p<0.01$. Again, the exceptions are MAE results for Twitter in the NYT dataset. Although news values achieve the

Table 4: Regression results comparing feature groups ($\mathcal{M}_\mathcal{N}$ = news values, $\mathcal{M}_\mathcal{S}$ = style, $\mathcal{M}_\mathcal{M}$ = metadata). Result in bold indicates improvement of $p<0.01$.

	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
$\mathcal{M}_\mathcal{M}$	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**

lowest performance of all groups, the correlation with Twitter and Facebook popularity is still between 0.14 and 0.2. It is especially noteworthy that style features, which are largely topic-independent, on their own achieve good performance (up to 0.25 correlation, and 0.7 MAE). This suggests that headline style is important to social media readers, independent of article content. This seems to follow previous research on online content popularity prediction, where various aspects of style were also found to have an impact on popularity (Tan, Lee, and Pang 2014). Metadata (especially category) achieves good results, in particular for the Guardian dataset, suggesting that topic and genre of the article play a significant role for readers. Although metadata adds to the prediction performance, it should be noted that this aspect of news articles is usually not controlled by the writer (i.e. one cannot easily change the genre or the topic). On the other hand, most news values and style features can be freely edited, in order to reach higher popularity.

Differences between Twitter and Facebook. Perhaps due to a more skewed distribution, Facebook has higher errors. For correlations, Twitter performs higher with Guardian data, while it is the opposite for NYT. Different demographics of news readers on Facebook and Twitter[11] might contribute to this, which calls for further work that takes into account user demographics.

Differences between news sources. A key aspect of our work is the source-internal evaluation, which has not been done for this task before. Indeed, the performance is better on Guardian data than the NYT. This points at further work with other news outlets and genres (e.g. tabloids).

Computation cost vs. performance. When comparing against the best-performing baseline ($\mathcal{M}_\mathcal{A}$), our full model achieves significant improvement (cf. Table 2). The difference is much more noticeable when considering only features extracted directly from headline text (cf. Table 3). While the overall performance using all available features only slightly improves over the state-of-the-art, the model that uses features that can be more readily edited by the headline author (and possibly increase its popularity) shows considerable improvement.

[11] http://pewrsr.ch/27TOfhz

Conclusion

News headlines play a crucial role on social media. In a novel task to predict the social media popularity of news articles using headline-derived features, we improved significantly over several baselines. Features extracted from headline text (which usually can be edited by the headline author) were shown to have impact on the prediction performance when considered on their own. This suggests that traditional editorial judgments about newsworthiness and insights from NLP research on style are applicable to predicting headline popularity on social media. Our feature extraction methods are generic and can be repeated across different news outlets and genres. The results of the prediction model depend on the news source; further work can include performance comparison across different news outlets and online content.

We are currently refining the prediction model taking into account user demographics and integrating world knowledge. Firstly, we are considering user location (country of residence) to improve the Proximity feature. Secondly, to improve Prominence (our best-correlated feature) we are incorporating world knowledge from Wikidata to relate entity significance to the user's location.

Acknowledgments

This work was supported by a Doctoral Training Grant from the EPSRC, UK. Data collection and storage comply with EPSRC data management policies. The dataset is available at https://doi.org/10.5518/174.

References

Arapakis, I.; Cambazoglu, B. B.; and Lalmas, M. 2014. On the feasibility of predicting news popularity at cold start. In *SocInfo*, 290–299.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*.

Bandari, R.; Asur, S.; and Huberman, B. A. 2012. The pulse of news in social media: Forecasting popularity. In *ICWSM*.

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

Castillo, C.; El-Haddad, M.; Pfeffer, J.; and Stempeck, M. 2014. Characterizing the life cycle of online news stories using social media reactions. In *CSCW*.

Holsanova, J.; Rahm, H.; and Holmqvist, K. 2006. Entry points and reading paths on newspaper spreads: comparing a semiotic analysis with eye-tracking measurements. *Visual communication* 5(1):65–93.

Lakkaraju, H.; McAuley, J. J.; and Leskovec, J. 2013. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In *ICWSM*.

Piotrkowicz, A.; Dimitrova, V. G.; and Markert, K. 2017. Automatic extraction of news values from headline text. In *Proceedings of the EACL 2017 Student Research Workshop*.

Tan, C.; Lee, L.; and Pang, B. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. In *ACL*.