Automatic Extraction of News Values from Headline Text

Alicja Piotrkowicz
School of Computing
University of Leeds
scap@leeds.ac.uk

Vania Dimitrova
School of Computing
University of Leeds
V.G.Dimitrova@leeds.ac.uk

Katja Markert
Institut für Computerlinguistik
Universität Heidelberg
markert@cl.uni-heidelberg.de

Abstract

Headlines play a crucial role in attracting audiences' attention to online artefacts (e.g. news articles, videos, blogs). The ability to carry out an automatic, large-scale analysis of headlines is critical to facilitate the selection and prioritisation of a large volume of digital content. In journalism studies news content has been extensively studied using manually annotated news values – factors used implicitly and explicitly when making decisions on the selection and prioritisation of news items. This paper presents the first attempt at a fully automatic extraction of news values from headline text. The news values extraction methods are applied on a large headlines corpus collected from *The Guardian*, and evaluated by comparing it with a manually annotated gold standard. A crowdsourcing survey indicates that news values affect people's decisions to click on a headline, supporting the need for an automatic news values detection.

1 Introduction

In this digital age, where "the widening gap between limitless media and limited attention makes it a challenge for anything to attract an audience" (Webster, 2014), headlines play a special role. Their main function is to draw attention and act as the visual entry point to online digital content (Leckner, 2012). This is intensified on social media, where in cases of indirect engagement (e.g. with retweeted news articles) headlines are often the only visible part of the main content. Liu (2005) found that compared to print media, digital readers spend more time browsing, scanning, and keyword spotting. Various studies conducted by Chartbeat found that 38% of users leave a website immediately after accessing it[1], and that an average reader will spend only 15 seconds on a website[2]. An American Press Institute study found that roughly six in ten people acknowledge that they are "headline-gazers" checking only the headline and not reading the full article [3].

Therefore, automatic processing of headlines is needed to facilitate the selection and prioritisation of large volumes of digital content. This has been studied in the journalism field by considering news values. These are aspects of an event determining whether and to what extent it is reported, therefore guiding editorial selection. Recent journalism research (O'Neill and Harcup, 2009, p.171) suggests that news values can also be applied to the audience reception perspective, thus helping to analyse what attracts audiences to certain headlines.

The automatic extraction of news values from headlines can be a central tool for a range of applications. Automatically extracted news values scores can be correlated with online attention metrics, such as pageviews, to investigate which headline aspects influence online popularity. They can play a key role in content-based recommender systems, especially when a user model is not available (the so-called 'cold start' problem). Headline newsworthiness insights can be incorporated into online content publishing, e.g. YouTube[4] to guide authors on how to compose the headline text to attract audiences' attention. Furthermore, digital humanities researchers can conduct large-scale comparisons of news values across digital outlet types, genres, demographics, etc.

Despite the importance of headline news values, there are no automatic computational means to extract them from headline text. This requires advanced text processing to compute appropriate

[1] http://slate.me/1cJ7b5C
[2] http://yhoo.it/2cEQMVC
[3] http://bit.ly/21LwfS5
[4] https://www.youtube.com/

Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 64–74, Valencia, Spain, April 3-7 2017.

features that can be related to news values. It makes for a challenging problem, because news values often involve tacit knowledge. There are no precise definitions of news values which can be used for automatic text processing, which is further aggravated by the nature of headline text. Critically, there are no studies to inform how to associate news values with various features that can be automatically extracted from headline text.

To address these challenges we utilise state-of-the-art techniques to develop a method for automatic extraction of news values from headline text. Our solution includes several NLP methods, such as wikification, sentiment analysis, and language modeling. We further combine them with other AI methods, including a burst detection algorithm to propose new techniques for estimating entities' prominence. The approach is applied and evaluated on a large corpus of news headlines from a prominent news source – *The Guardian*.

Focusing on headline news values, the paper presents a new perspective on processing digital content and contributes to text analytics by: (i) providing the first computational method for a fully automatic extraction of news values from headlines which combines relevant NLP techniques; (ii) evaluating the news values feature engineering by applying the computational method to a large corpus of news headlines and comparing the automatic annotation to a gold standard developed for this task, (iii) confirming through a user crowdsourcing study that people's choices to click on news items are influenced by news values in the headlines, indicating the significance of automatic news values detection.

2 Related Work

Headlines are gaining ground in the NLP community as a text type to be studied separately. This follows research suggesting that headlines can function autonomously from the full text. According to Dor (2003) the reader receives "the best deal in reading the headline itself". Empirical studies seem to support this – Gabielkov et al. (2016) found that 59% of shared news content on Twitter is not clicked on, i.e. has not been read before being shared. This makes headlines key for sharing content on social media. In the journalism community, the importance of headlines has already been acknowledged. For example, Althaus et al. (2001) looked at substitutes for full article text including headlines and their impact on content analysis. Tenenboim and Cohen (2013) conducted a study on the effect of headline content on clicking and commenting. However, these efforts included a manual annotation, which limited their scope. More recently, NLP researchers also focused on headlines, including headline generation (Gatti et al., 2016) and keyword selection for popularising content (Szymanski et al., 2016). We add to this ongoing NLP research by proposing news values to analyse headlines.

News values originated in the journalism studies field with the work by Galtung and Ruge (1965). Since then a variety of taxonomies of news values have been proposed: Bell (1991), Harcup and O'Neill (2001), Johnson-Cartee (2005) and Bednarek and Caple (2012). Regardless of differences in granularity and definitions, there is a considerable overlap between all these taxonomies. This allows us to select the news values which are most frequently mentioned and most relevant to headline text. These include: prominence, sentiment, superlativeness, proximity, surprise, and uniqueness. We offer a systematic and fully replicable method of an automatic extraction of these news values from headlines. Furthermore, we show that these news values influence people's decisions to click on a headline.

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 journalism studies have focused on aspects that are readily available through article metadata (e.g. topics in Bastos (2014)). 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. For example, Potts et al. (2015) manually choose news values indicators from a preprocessed corpus; moreover, the approach relies on keywords and is topic-dependent. This paper presents the first attempt at a fully automatic and topic-independent extraction of news values which is applied and validated on headlines from a 'broadsheet' news source. Our news values detection is largely not news-specific

and can be extended to titles in other genres.

From an NLP perspective headlines pose an engineering challenge. This includes linguistic aspects like unusual use of tenses (Chovanec, 2014) and deliberate ambiguity (Brône and Coulson, 2010). There are also some domain-specific phenomena like click-baiting (Blom and Hansen, 2015). Headlines are typically short, which limits the amount of context that many NLP tools rely on. While feature engineering from headlines is less studied, there are research efforts that specifically address short texts. Tweets have attracted considerable attention, leading to the development of some Twitter-specific tools (e.g. TweetNLP[5]). Tan et al. (2014) is an example of feature engineering from tweets that looks specifically at wording and its effect on popularity. Another example of a text closely related to headlines are online content titles, e.g. image titles on Reddit (Lakkaraju et al., 2013). Many approaches include features like ratios for various parts of speech, sentiment, and similarity to a language model. However, they 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)). For each news value we either re-implement the most appropriate state-of-the-art methods, or implement new techniques that work well with headlines.

3 Extraction of News Values

We present feature engineering methods for six news values. These six were selected, because they occur frequently in news values taxonomies (cf. Section 2). The feature computation methods are summarised in Table 2. Although our goal is a generic framework, we are inspired by research in the news domain. Consequently, the features are informed by news values related to news content.

Preprocessing. All headlines are part-of-speech tagged (Stanford POS Tagger (Toutanova et al., 2003)) and parsed (Stanford Parser (Klein and Manning, 2003)). Wikification (a method of linking keywords in text to relevant Wikipedia pages; e.g. Mihalcea and Csomai (2007)) is used to identify entities in the text. Headlines are wikified using the TagMe API[6], a tool meant for short texts, making it suitable for headlines.

[5]http://www.cs.cmu.edu/ ark/TweetNLP/

[6]http://tagme.di.unipi.it/

Notation. We see the headline H as a set of tokens obtained from the POS tagger. We denote the set of content words in H as C and the set of entities in H as E (cf. Table 1).

Table 1: Preprocessing: H (set of tokens), C (set of content words), E (set of wikified entities)

"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 }

NV1: Prominence. Reference to prominent entities (elite nations and people (Galtung and Ruge, 1965), and more recently celebrities (Harcup and O'Neill, 2001)) is one of the key news values.

We approximate prominence as the amount of online attention an entity gets. As online prominence varies with time we consider long-term vs. recent prominence and burstiness. We extend previous work by using wikification for obtaining entities and considering their burstiness.

For an entity e, we denote as $pageviews_{e,d_{-m},d_{-n}}$ the median number of Wikipedia daily page views[7] 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 recent prominence on the day before publication ($pageviews_{e,d_{-1},d_{-1}}$).[8] For a news-centric perspective of prominence, we also calculate the sum of e's mentions in the news source headlines in the week before publication day, denoted as $newsmentions_{e,d_{-7},d_{-1}}$.

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). 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 1). We use entity bursts in two ways. Firstly, burstiness indicates the number of days that e was in a burst over a year ($daysburst_{e,d-365,d_{-1}}$). Sec-

[7]http://dumps.wikimedia.org/other/pagecounts-ez/

[8]We found the previous day's prominence to be closest to the actual on-the-day prominence.

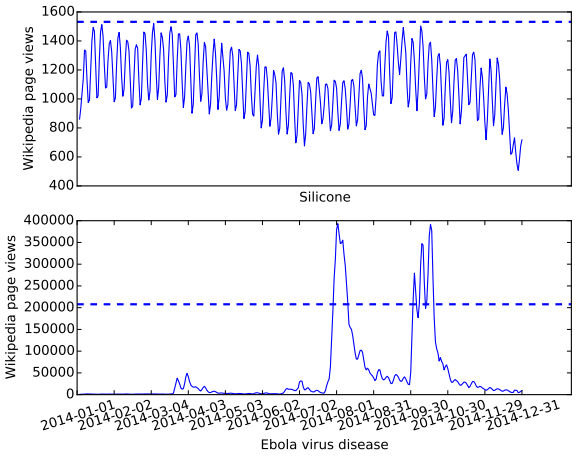

Figure 1: Time series plots of Wikipedia page views moving averages (MA) for two entities: non-bursty SILICONE (top) and bursty EBOLA VIRUS DISEASE (bottom). The dashed line shows the burst cut-off line.

ondly, 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). We are the first to consider burstiness for popularity prediction.

Algorithm 1 Burst detection algorithm adapted from Vlachos et al. (2004). Following experimentation, moving average was set 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 = $d_i | MA_e(i) > cutoff$

As a headline can have multiple entities, all prominence measures are aggregated via summation over all entities in H (see Table 2).

NV2: Sentiment. This refers to sentiment-charged events (Johnson-Cartee, 2005) and using sentiment-charged language (Bednarek and Caple, 2012). Features relating to sentiment and emotionality have been shown to influence a news article's virality (Berger and Milkman, 2012). However, this effect has not been studied for headlines.

As direct measures of sentiment, we combine SentiWordNet (Baccianella et al., 2010) positivity and negativity scores of content words, and calculate sentiment and polarity scores following Kucuktunc et al. (2012). Sentiment can also be indirect. Firstly, a word may be in itself objective, but carry a negative connotation (e.g. *scream*). We therefore measure the percentage of content words in a headline with a positive or negative connotation (using a connotations lexicon (Feng et al., 2013)). Secondly, we measure the percentage of biased content words (using a bias lexicon (Recasens et al., 2013)). 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 reading.

NV3: Superlativeness. The size (Johnson-Cartee, 2005, p.128), or magnitude (Harcup and O'Neill, 2001) of an event is considered to influence news selection.

We focus on explicit linguistic indicators of event size: comparatives and superlatives (indicated by part-of-speech tags), and amplifiers (indicated with intensifiers and downtoners). For the latter, we combine the lists in Quirk et al. (1985) and Biber (1991), obtaining wordlists of 248 intensifiers and 39 downtoners.

NV4: Proximity. This news value has been interpreted as both geographical (Johnson-Cartee, 2005, p.128) and cultural proximity (Galtung and Ruge, 1965) of the event to the news source or the reader (Caple and Bednarek, 2013).

Following an assumption that readers from the country of a news outlet constitute the main part of its readership, we focus on geographic proximity to the news source. We use a binary feature that indicates whether a headline refers to an entity that is geographically close to the news source, and manually create a wordlist including names for the country, regions, capital city (17 UK-related terms in total). We then look for matches in the headline text ("*London* smog warning as Saharan sand sweeps southern *England*") or the Wikipedia categories of each entity supplied in the TagMe output (category POSTAL SYSTEM OF THE UNITED KINGDOM for headline "Undervaluing *Royal Mail* shares cost taxpayers £750m in one day").

NV5: Surprise. Events which involve "surprise and/or contrast" (Harcup and O'Neill, 2001) make news. 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.

We target explicit surprise by calculating the commonness of phrases in headlines with reference to a large corpus. We first extract phrases of following types: SUBJ-V, V-OBJ, ADV-V, ADJ-

Table 2: Feature implementations and statistics on *The Guardian*. Notation is in Table 1. Measures: median and maximum values, prevalence (proportion of non-zero scores), and the Kruskall-Wallis test comparing the manual gold standard to automatic extraction (* $p<0.05$, ** $p<0.01$, *** $p<0.001$).

	Feature name	Implementation	Median	Max	Prevalence	KW
NV1	number of entities	$\|E\|$	1	8	79%	***
	Wikipedia current burst size	$\sum_{e \in E} daysburst_{e,d_{-1},d_{-1}} \times \frac{pageviews(e,d_{-1},d_{-1}) - \text{mean}(MA_e)}{\text{SD}(MA_e)}$	0	57.16	12%	0.2
	Wikipedia burstiness	$\sum_{e \in E} daysburst_{e,d_{-365},d_{-1}}$	21	156	78%	***
	Wikipedia long-term prominence	$\sum_{e \in E} pageviews_{e,d_{-365},d_{-1}}$	1,342	125,757	79%	***
	Wikipedia day-before prominence	$\sum_{e \in E} pageviews_{e,d_{-1},d_{-1}}$	1,642	1,031,722	78%	***
	News source recent prominence	$\sum_{e \in E} newsmentions_{e,d_{-7},d_{-1}}$	0	122	50%	**
NV2	sentiment	$max\_positivity - max\_negativity - 2$	-2	-1	100%	0.1
	polarity	$max\_positivity + max\_negativity$	0.5	1.88	79%	**
	connotations	$\frac{\text{\# content words with positive or negative connotations}}{\|C\|}$	0.34	1	92%	0.2
	bias	$\frac{\text{\# biased content words}}{\|C\|}$	0.13	1	61%	*
NV3	comparative/superlative	$\frac{\text{\# words with JJR\|JJS\|RBR\|RBS POS tag}}{\|C\|}$	0	1	7%	***
	intensifiers	$\frac{\text{\# intensifiers}}{\|H\|}$	0	0.34	10%	***
	downtoners	$\frac{\text{\# downtoners}}{\|H\|}$	0	0.29	4%	0.2
NV4	proximity	1 if explicit reference to UK in H or in Wikipedia category tags, else 0	0	1	35%	***
NV5	surprise	$min\text{LL}_p$ where LL_p is the log-likelihood for a phrase in H	4.15	2,726,186	100%	*
NV6	uniqueness	$max_{t \in d-72hr}$ cosine similarity$(H, pastH_t)$	0	0.83	13%	*

N, N-N; and generate a regular expression with their inflected forms (e.g. *man drinks* → *man drinks|drank|drinking*). For each regexp we obtain a count from a Wikipedia corpus[9] and sum the counts for each phrase and calculate its log-likelihood (LL). The feature value is given the lowest LL in the headline (as we are looking for the most surprising phrase)[10].

NV6: Uniqueness. News has to be new – "any new comment or circumstance [...] adds to the debate" (Conley and Lamble, 2006). An analysis of several storylines in the headlines corpus showed that of two very similar headlines, the latter tends to be less popular ("Ferry disaster: South Korean prime minister resigns" was more popular than the later "South Korean prime minister resigns over ferry sinking").

For a headline H we select past headlines from 72 hours before H's publication and which have at least one TagMe entity overlapping or neither has any entities[11]. 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.

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

[10] We experimented with other corpora and metrics and found Wikipedia and log-likelihood to give best results.

[11] Entity overlap helps with ensuring that the headlines are part of the same storyline; including headlines with no entities ensures more coverage. Collecting headlines from previous 72 hours works better than other cutoff points.

4 Application and Evaluation

We applied the feature extraction methods on a corpus of headlines from *The Guardian*, a major British newspaper. This provides a wide coverage of various topics and genres, allowing a good exploration of news values. The automatic extraction of news values was compared to a manually annotated gold standard.

Headline corpus. The headlines corpus was built using the Guardian Content API[12]. We downloaded all headlines published during April 2014, yielding a corpus of 11,980 headlines.

Automatic annotation. Feature values were calculated for each headline. Statistics for the extracted features in *The Guardian* corpus are reported in Table 2 (Median, Max, Prevalence).

Manually annotated gold standard. For each news value we selected 20 headlines from the headlines corpus. In order to use the clearest examples for a more accurate annotation, we randomly selected 10 headlines from the top quartile values and 10 from the bottom quartile. For news values that are split into multiple features (NV1:Prominence, NV2:Sentiment, NV3:Superlativeness), the feature group vectors were ordered to obtain quartiles. Overall, a total of 120 headlines were selected for manual annotation. Three expert annotators, PhD students in linguistics, annotated each headline as positive or

[12] http://www.theguardian.com/open-platform

negative (Y/N) for the first five news values (cf. Table 3). For NV6:Uniqueness, annotators were presented with 20 headlines from the corpus and further 20 past headlines with highest and lowest headline uniqueness scores (which were randomly sampled). The annotators indicated whether any of the past headlines were very similar (i.e. highly related) to a given headline.

Inter-annotator agreement. The inter-annotator agreement was calculated using Fleiss's Kappa. It ranges from substantial for NV1:Prominence (.76) and NV6:Uniqueness (.73), through moderate for NV3:Superlativeness (.43), NV5:Surprise (.48), and NV4:Proximity (.55), to fair for NV2:Sentiment (.22). The annotators remarked that sometimes they chose 'on instinct' and their responses might vary from day to day. This highlights the challenge of an automatic detection of news values, as news values are somehow tacitly understood. The annotators' judgments were aggregated using a majority vote, creating the gold standard.

Comparison with gold standard. We calculated pairwise comparisons between each feature and the relevant manual label (e.g. number of entities and Prominence, bias and Sentiment). The Kruskal-Wallis test was used to determine whether the differences in feature values for the two manual annotation labels (Y/N) were significant (cf. column KW in Table 2). 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 experts. The findings of the evaluation are discussed below.

5 Discussion of Feature Extraction

We use a news corpus that is representative of a wide range of news publications under the umbrella of 'broadsheet' (as opposed to tabloid newspapers which differ in style and tone). *The Guardian* corpus is a freely available resource, allowing replication of methods and study findings. While the evaluation of feature extraction is conducted over one corpus, we also applied this approach to another publicly available 'broadsheet' corpus – *New York Times* (cf. Appendix A). We will discuss below the findings from *The Guardian* evaluation study, and will refer to feature extraction outputs from *New York Times* to illustrate feature behaviour on two corpora.

NV1: Prominence is one of the most prevalent news values and our approach using wikification proves very reliable. 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 (max. 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 the computation of prominence. Wikipedia current burst size is a rare feature (12% in *The Guardian* and 10% in NYT), because capturing an entity in a burst is uncommon, since bursts do not apply to all entities and do not happen frequently.

The IAA for Prominence is the highest (κ=.76) and nearly all features reach $p<0.001$ 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 presents a new way of looking at Prominence. While burstiness (i.e. how many times in a year an entity had pageviews 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 correlated with the gold standard.

NV2: Sentiment is among the most challenging news values to implement, since it is not typical for broadsheets and sentiment-charged language in headlines does not always accurately reflect the true sentiment or emotion. Headlines in broadsheet newspapers tend to be quite neutral (median sentiment = -2; median polarity = 0.5). This is also the case for the *New York Times* (sentiment = -2; polarity = 0). However, most headlines contain at least one connotated or biased word (connotations prevalence = 92%, bias prevalence = 61%; slightly lower in NYT: 78% and 51%).

The IAA 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. Furthermore, while manual annotation for one aspect of Sentiment like positivity/negativity can achieve substantial agreement (.76 agreement between experts in Snow et

Table 3: Examples of annotated headlines. Y/N: majority vote manual annotation. Below: automatically extracted values aggregated via summation by feature group (cf. Table 2 for feature value ranges).

#	Headline	Prominence	Sentiment	Superlativeness	Proximity	Surprise
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

al. (2008)), our definition of Sentiment is broader. The annotators pointed out an interesting characteristic of expressing Sentiment. On one hand, there were highly evocative headlines that describe some tragic news events (+sentiment, +emotion). On the other hand, some headlines use sentiment-charged language, but were not evocative to the same extent (+sentiment, -emotion). For example, *comedy* (E5 in Table 3) has positive sentiment, but does not evoke positive emotion. When compared to the manual annotations, two 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?"). Disentangling sentiment and emotion might paint a clearer picture.

NV3: Superlativeness is rare, but reliably extracted. It is the least prevalent news value (between 4-10%; between 3-6% in NYT). The median values are also all zero. Our narrower definition of could be the reason, however we decided to focus on explicit linguistic indicators of event size (e.g. *very*, *hardly*) to keep the implementation topic-independent and more easily generalisable.

The IAA was moderate (κ=.43). Two out of three features were significant at $p<0.001$. This confirms that our approach that relies on POS tags and wordlists does capture 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%), 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 Superlativeness features (comparative/superlative and intensifiers) can be reliably used for headlines.

NV4: Proximity is not frequent, but our approach using a wordlist and Wikipedia categories proves very reliable. This news values occurs in 35% of headlines. This is not surprising, considering that *The Guardian* has a global audience, so the majority of news is not UK-specific (prevalence in NYT is similar at 32%).

The IAA is moderate (κ=.55). The feature reaches significance at $p<0.001$, so our method of capturing Proximity 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 NER/wikification tools. In some cases an entity might be missed (cf. E1 in Table 3, where *EE/Orange* was missed and consequently both Prominence and Proximity scores are zero). It is important to note that Proximity covers both geographic and cultural proximity. Our annotators were UK residents, familiar with *The Guardian*, but demographics of the reader will probably influence their familiarity with some entities. In our future work we will include some demographics data to deepen the implementation for Proximity.

NV5: Surprise is difficult to implement due to peculiarities of headline text, but our approach which targets surprising phrasing using a Wikipedia-based language model does capture it. The median log-likelihood for this features is relatively low (4.15; 4.04 for NYT), 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). When using a corpus which is not specifically for headlines (we used Wikipedia), the log-likelihood will tend to be lower.

The IAA was moderate at κ=.48 and the feature is significant ($p<0.05$). This shows that using a count-based method captures 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.

NV6: Uniqueness, or rather a lack of it, is fairly rare, but our implementation reliably identities such instances. The prevalence is quite low (15%; but slightly higher at 34% in *New York Times*), which follows the basic journalistic principle that news have to be novel.

IAA was substantial with κ=.73 and 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 media like video).

Overall, 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 our approach successfully identified and quantified at least some aspects of every news value.

The study also indicated open issues requiring further investigation. Firstly, the findings highlight the importance of world knowledge when analysing headlines. For example, for the well-established NLP topic like sentiment analysis, we find that although purely linguistic methods can capture most phenomena in headlines, they fall short to recognise sentiment within entities (e.g. Ebola). Similarly, a more generic approach for Proximity would require world knowledge to detect that an entity is related to the reader's location. We are addressing this in our future work. Secondly, it will be interesting to explore how the proposed methods can be applied to other types of news sources (e.g. tabloids) and to genres other than news. With the exception of news source prominence and uniqueness, our features are not news-specific. Titles for other types of digital content (blogs, videos) also include prominent entities, sentiment or intensifiers. News values detection offers a new perspective for their analysis. Thirdly, our methods can be adapted to other languages, provided that certain NLP resources exist (POS tagger, NER, sentiment lexicon). This would enable large-scale analyses of headlines along multiple axes, like language and genre.

6 Do News Values Influence People's Choice of Headlines?

To show the importance of the automatic news value extraction for a range of applications (cf. Section 1), we examined whether news values matter for general audiences. This was explored with a crowdsourcing study.

Survey content. The survey consisted of five short sections for news values NV1 to NV5 (NV6:Uniqueness 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 section participants were presented with a short definition and several examples. Then they were asked the following: *"I personally consider this news value when clicking on headlines"* and given five Likert scale responses (cf. Figure 2). Standard demographics information (age, gender, country of residence, native language, news reading habits) was collected.

Participants. The crowdsourcing platform CrowdFlower was used to recruit participants for the survey, allowing 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, 96 were recorded as complete. 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: 48 participants were 34 or younger and 23 were 35 or older; 17 were female, 54 were male; 30 were native English speakers and 41 were non-native English speakers; 44 participants read news daily, 27 weekly.

Results and discussion. Results are presented in Fig. 2. The overall impact that news values have on survey participants has been indicated as very positive. NV1:Prominence, NV4:Proximity,

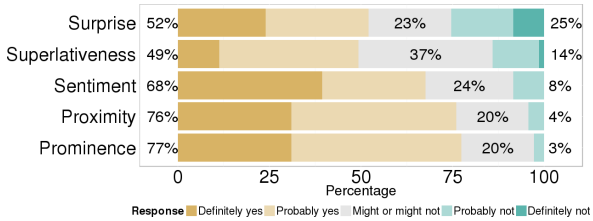

Figure 2: Survey results to the question *"I personally consider this news value when clicking on headlines"* (N=71). Percentages show aggregated positive, neutral, and negative responses.

and NV2:Sentiment had the highest proportions of positive answers (77%, 76%, and 68%, respectively). This follows the journalism studies literature, where these three news values attract perhaps the most focus. Comparison with the gold standard confirmed that our implementation for NV1:Prominence and NV4:Proximity reflects the experts' judgments. Since this survey highlighted the role of Sentiment, we are motivated to develop it further to capture its full extent. NV3:Superlativeness had the most neutral responses (37%). On one hand, this could be because this news value is slightly more difficult to understand[13]. On the other hand, Superlativeness might have been deemed to play a lesser role, since its main function is more supportive (to embellish or diminish content). Finally, NV5:Surprise had the most negative responses (25%). This might be because surprising headlines could be perceived as less informative, or more ambiguous. As people often read only headlines to get their news (Gabielkov et al., 2016), surprise would not support the headlines' function as summaries.

Overall, results of this survey highlight the importance of news values in headlines. We also found that news values play a role for both native and non-native speakers of English (our sample has roughly equal numbers of both). This is important, since most major news outlets nowadays have a more global reach.

7 Conclusions and Future Work

The work presented here is the first step in a larger project to predict the popularity of news articles using headlines. Our focus on headlines is motivated by their role in the everyday online experience, characterised by limited audience attention and the frequent use of social media websites.

We proposed an automatic extraction method for *news values*, which have been posited in journalism studies and offer a new perspective on characterising digital content. We broke novel ground by developing fully automatic and topic-independent methods for identifying news values in headlines. An evaluation using manual annotations shows that for all news values the output of the automatic extraction corresponds to the gold standard. The results from a crowdsourced survey indicated that news values influence people's decisions to click on a headline. This supports the wider adoption of the automatic method of analysing headlines in a range of applications concerning human choices (e.g. prediction models, recommender systems, intelligent assistants).

Our current and future work includes several stages. Firstly, we have collected a second corpus (*New York Times*) to apply our news values extraction methods. Secondly, the extracted news values scores are being correlated with popularity of headlines on social media and applied in a popularity prediction model using machine learning methods. The results from the manual annotations and the crowdsourced survey will also be used to inform the weights of features in the prediction model. Furthermore, another survey will target the direct engagement with headlines (i.e. whether a reader would click the headline) and compare it to the social media popularity metrics we have already collected. Finally, using both data from the crowdsourced surveys and publicly available Twitter data we will look at whether demographics, in particular the country of residence, have impact on the news values of Prominence and Proximity. We will use the data on the entities we identified from knowledge bases like *Wikidata* and *BabelNet* to enrich the implementations of these news values.

Acknowledgments

This work was supported by a Doctoral Training Grant from the Engineering and Physical Sciences Research Council. Data collection and storage comply with EPSRC data management policies. The dataset is available at `http://doi.org/10.5518/147`.

We would also like to thank our expert annotators for their work and feedback.

[13] 57% of native English speakers judged Superlativeness positively compared to 44% of non-native speakers.

References

Scott L. Althaus, Jill A. Edy, and Patricia F. Phalen. 2001. Using substitutes for full-text news stories in content analysis: Which text is best? *American Journal of Political Science*, 45(3):pp. 707–723.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. 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 conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Marco Toledo Bastos. 2014. Shares, pins, and tweets: News readership from daily papers to social media. *Journalism Studies*, pages 1–21.

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

Monika Bednarek and Helen Caple. 2014. 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.

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

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

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

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

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

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

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

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

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

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. 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, Association for Computational Linguistics.

Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. 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.

Johan Galtung and Mari Holmboe Ruge. 1965. 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.

Lorenzo Gatti, Gözde Özdal, Marco Guerini, Oliviero Stock, and Carlo Strapparava. 2016. Automatic creation of flexible catchy headlines. In *Natural Language Processing meets Journalism Workshop*. IJCAI.

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

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

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

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

Onur Kucuktunc, Berkant Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu. 2012. 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.

Himabindu Lakkaraju, Julian J. McAuley, and Jure Leskovec. 2013. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 311–320

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

Ziming Liu. 2005. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of documentation*, 61(6):700–712.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, ACM.

Deirdre O'Neill and Tony Harcup. 2009. News values and selectivity. *The Handbook of Journalism Studies*, pages 161–174.

Amanda Potts, Monika Bednarek, and Helen Caple. 2015. How can computer-based methods help researchers to investigate news values in large datasets? *Discourse & Communication*, 9(2):149–172.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Longman.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. 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, Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. 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, Association for Computational Linguistics.

Terrence Szymanski, Claudia Orellana-Rodriguez, and Mark T. Keane. 2016. 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*. IJCAI.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland, June. Association for Computational Linguistics.

Ori Tenenboim and Akiba A. Cohen. 2013. What prompts users to click and comment: A longitudinal study of online news. *Journalism*, 16(2):198–217.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. 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.

Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. 2004. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 2942.

James G. Webster. 2014. *The marketplace of attention: How audiences take shape in a digital age*. MIT Press.

A Supplementary Material: Feature Extraction on *New York Times*

Table 4: Feature extraction statistics on *New York Times* corpus. Notation is explained in Table 1. Reported measures: median and maximum values, prevalence (proportion of non-zero scores). WP=Wikipedia.

Feature name	Median	Max	Prevalence
NV1: Prominence			
Number of entities	1	4	100%
WP current burst size	0	57.18	10%
WP burstiness	15	166	66%
WP long-term prominence	626	65,327	66%
WP day-before prominence	773	467,458	66%
News source recent prominence	0	70	32%
NV2: Sentiment			
Sentiment	-2	-1	100%
Polarity	0	1.88	43%
Connotations	0.25	1	78%
Bias	0.11	1	51%
NV3: Superlativeness			
Comparative/superlative	0	1	3%
Intensifiers	0	0.33	6%
Downtoners	0	0.33	3%
NV4: Proximity			
Proximity	0	1	32%
NV5: Surprise			
Surprise	4.04	2,724,886	100%
NV6: Uniqueness			
Uniqueness	0	1	34%