DIGITAL JOURNALISM
https://doi.org/10.1080/21670811.2021.1984266

Routledge
Taylor & Francis Group

ARTICLE

Anticipating Attention: On the Predictability of News Headline Tests

Nick Hagar[a], Nicholas Diakopoulos[a] and Burton DeWilde[b]*

[a]Communication Studies, Northwestern University, Evanston, IL, USA; [b]Chartbeat, Inc, New York, NY, USA

ABSTRACT

Headlines play an important role in both news audiences' attention decisions online and in news organizations' efforts to attract that attention. A large body of research focuses on developing generally applicable heuristics for more effective headline writing. In this work, we measure the importance of a number of theoretically motivated textual features to headline performance. Using a corpus of hundreds of thousands of headline A/B tests run by hundreds of news publishers, we develop and evaluate a machine-learned model to predict headline testing outcomes. We find that the model exhibits modest performance above baseline and further estimate an empirical upper bound for such content-based prediction in this domain, indicating an important role for non-content-based factors in test outcomes. Together, these results suggest that any particular headline writing approach has only a marginal impact, and that understanding reader behavior and headline context are key to predicting news attention decisions.

KEYWORDS

Digital journalism; headline writing; news attention; text analysis; news values; news production; computational methods

Introduction

Newsrooms increasingly optimize their output for audience attention in both coverage areas and presentation (Anderson 2011; Fürst 2020; Petre 2015) hoping to ensure their viability in an attention-dependent digital media marketplace (Webster 2016). But it's hard to judge how much a news producer's decisions actually matter in attracting audience attention. To be sure, news audiences are influenced by the articles, modes of presentation, and writing styles that news organizations employ (Jungherr, Posegga, and An 2019; Kuiken et al. 2017). However, readers also rely on their personal preferences and backgrounds when composing their news diets (Garrett and Stroud 2014; Lamberson and Soroka 2018). The interactions between these influences are difficult to untangle (Kessler and Engelmann 2019).

CONTACT Nicholas Diakopoulos nad@northwestern.edu
*Chartbeat, Inc, New York, NY, USA is prior affiliation during period in which study was conducted.

Supplemental data for this article can be accessed at https://doi.org/10.1080/21670811.2021.1984266.

One focus of audience-facing optimization is the headline, the first (and often only) piece of text a reader might see from a news article. Headlines must summarize articles' contents and entice readers to click; a goal of headline optimization is to accomplish both with a concise and attractive writing style (Dor 2003). Many studies have attempted to detail effective, attention-grabbing headline writing strategies (Kim et al. 2016; Kuiken et al. 2017; Rayson 2017). A/B testing allows journalists and editors to write multiple variations of an article's headline, display those variations to different members of the audience, then select the best-performing version (Hagar and Diakopoulos 2019).

Many of these efforts stop short of addressing broader questions about how news attention works online. By considering headline writing in isolation, they tend to overlook the role of audience preferences, attitudes, and habits (Kormelink and Meijer 2018). Predicting the performance of any piece of digital content solely based on its composition is a known challenge, in tension with the idea that general heuristics for headline writing can consistently improve performance (Arapakis, Cambazoglu, and Lalmas 2017; Martin et al. 2016).

This work interrogates the relationship between headline writing style, audience factors, and performance. We focus on headlines that appear on news publisher home and section pages, in the context of A/B tests, using a large-scale, real-world dataset. First, we examine the extent to which a headline's textual features can predict its performance. Second, within those bounds of predictability, we demonstrate the relative contribution of content-based features to headline performance.

Our predictive model achieves modest success relative to our baselines while also demonstrating limits to content-based prediction. To further analyze the importance of writing style, we draw from a comprehensive range of theoretically and empirically motivated textual features. We show that, while several features' outcomes agree directionally with prior work, associations between feature usage and headline performance are weak overall. These findings suggest areas outside of a headline's composition, such as audience behavior and preferences, may warrant further study in understanding and predicting headline performance.

Background

In this section we consider related work on news audience attention and its relation to headlines. We also introduce background on the predictive approach we take in modeling that attention.

Audience Attention and Headlines

Perceptions, preferences, and beliefs shape what news sources a reader might seek out. Similarly, partisan preferences and attitudes toward news organizations impact what kinds of sources readers get exposed to (Flaxman, Goel, and Rao 2016; Kessler and Engelmann 2019). A reader's familiarity with a piece of news, and their interest in that news, can affect the extent to which it draws their attention (Lamberson and Soroka 2018). In addition, readers approach news with ingrained habits and routines

of consumption (Makhortykh et al. 2020). Individual preconceptions—attitudes and behaviors that are difficult for a news organization to alter—play a significant role in news attention decisions.

Journalists and editors pay special attention to writing effective headlines. In print, an effective headline summarizes or highlights an article's most interesting points (Ifantidou 2009). Digital headlines tend to prioritize drawing in an audience from across distribution platforms, since news organizations often depend on reader clicks for ad revenue and as the wide end of the funnel to subsequent subscription revenue (Anderson 2011; Petre 2015). That shift conflates what headline writers consider "good"—a value judgment traditionally based on professional standards of craft and ethics—and what gets clicked on the most, as seen in practices of algorithmically optimized content distribution such as headline A/B testing (Diakopoulos 2019; Hagar and Diakopoulos 2019; Ross 2017).

Rather than cultivating independent, editorial judgment, journalists may shift focus toward what the audience demands (Klinenberg 2005; Ross 2017). They in turn may place a greater emphasis on the kinds of coverage (e.g., soft news) and news values (e.g., proximity) that online audiences consider newsworthy (Trilling, Tolochko, and Burscher 2017). In terms of headline optimization, this can result in clickbait, which attempts to generate curiosity by implicitly referencing material in the article without revealing its details (Blom and Hansen 2015). These changes represent a shift in the specific qualities practitioners uphold as best practices in headline writing.

Using the direct measurement of audience response made possible by testing and other analytics tools, researchers and practitioners can more precisely evaluate the performance impact of writing strategies. We take this evaluation a step further by predicting headlines' performance from their contents, providing a view of how broadly writing approaches relate to the attraction and optimization of audience attention.

Headline Performance: From Explanation to Prediction

Using audience data, prior research has examined the effects of specific linguistic features on headline performance. Kuiken et al. (2017) measured headlines' click through rates in email newsletters and found a variety of textual features with a positive impact on headline performance, including average word length, lack of interrogatives, absence of quotes, use of personal or possessive pronouns, and presence of sentimental words. Industry researchers studied the performance of 100 million headlines on Facebook, albeit not all from news publishers, to extract specific phrases and emotions that elicited strong engagement (Rayson 2017). Kim et al. (2016) used news article click through rates from the Yahoo! homepage to evaluate the performance impact of various words and parts of speech. Much of this prior work operates through an explanatory lens: Using interpretable statistical approaches, studies attempt to demonstrate the extent to which proposed mechanisms are plausible drivers of headline success. In contrast, our research adopts a predictive approach, oriented toward developing models to predict unknown outcomes from previous observations.

Predictive modeling is complementary to more explanatory empirical work. First, predictive models help uncover new phenomena of interest. By identifying predictors that improve explanatory models, a predictive lens can enhance our empirical understanding of certain outcomes (Hindman 2015; Shmueli 2010). Second, prediction helps to establish the limits of what we might hope to understand about a phenomenon. As Tetlock and Gardner (2015) point out, knowing the limits of an outcome's predictability is itself valuable, in that it provides vital context to any accuracy measurements. Shmueli (2010) reinforces this notion, arguing that predictive models can help establish benchmarks for an outcome's potential explainability. Finally, predictive models help gauge the distance between theory and practice, testing how well proposed theoretical mechanisms apply in a given practical context (Shmueli 2010). As we elaborate further below, the features that we operationalize to support our predictive model are theoretically motivated, and the usefulness of those features in the model help to establish the external validity of those theoretical ideas in the specific context of news headline performance (Margolin 2019).

Predicting performance outcomes based on content alone is a known challenge. In many cases, predictive models require some early performance data from which to extrapolate (Szabo and Huberman 2010). Other factors, such as social influence on digital platforms, have been shown to affect performance outcomes more than content itself (Salganik, Dodds, and Watts 2006). This prediction difficulty also holds true for news headlines (Arapakis, Cambazoglu, and Lalmas 2017). Applying a predictive lens to any content-based performance outcome should seek to establish bounds on predictability. Our work addresses the following questions:

RQ1: To what extent can the text of a headline predict its performance?

RQ2: What is the relative importance of various content-based features to headline performance?

Data

Our data come from Chartbeat, a company that provides analytics services to digital publishers. This includes their Engaged Headline Testing system, which experimentally compares multiple versions of an article's headline to determine which is most effective at attracting readers. In a headline *test*, the system presents different readers with different headline *variants* for the same article and measures how many people click on each variant. As differences in performance emerge across variants, the system shows higher-performing variants to a larger portion of the site's audience.[1] Once the system is confident of a statistically significant difference in performance across variants, it marks a test as "converged" (described in more detail below).[2]

For each test, these data contain the text of the headline variants, each variant's associated performance in terms of clicks and impressions, and metadata about when and on which page a test was run. All tests in our sample take place on the homepages or section pages of news sites. The dataset represents direct comparisons of headline constructions, with real readers, across many news sites of different sizes and types. Because each headline variant is only compared to other variants within its test,

and each test corresponds to one article, these data are well-suited for isolating the way a headline is written from the contents of its corresponding article.

Data Filtering

The complete Chartbeat dataset represents 1,023,996 A/B headline tests with 2,662,572 headline variants, run across 1,314 web domains between April 1, 2015, and April 30, 2020. To limit our analysis to tests with clear results and clean data, we first filtered out certain classes of tests. Since the statistical models used in our natural language processing pipeline were trained on English corpora, we filtered out any tests with a headline not written in English. We did this by first removing all headline tests run on domains that were manually determined to publish articles in a language other than English, then by removing tests with at least one variant tagged as consisting of more than 20% non-English words. Next, we excluded any anomalous tests with a headline that had zero clicks recorded. In addition, we removed any tests for which the system did not reach statistical confidence about the winning variant[3], including those that were prematurely canceled by a user. Since publishers can run A/B tests on non-headline text (e.g. section tags or sub-headlines) we manually analyzed a random sample of 300 variants in our dataset, ranging from one to 48 words long. We found that most headlines fell between three and 30 words and excluded any test with a variant outside that range. This filtering pipeline retained 140,918 A/B headline tests and 334,976 headline variants across 293 domains.

Deriving Performance Metrics

To measure headline performance, we use a normalized version of click through rate, which we call *lift*. To calculate lift, we first computed the raw click through rate (CTR) for each variant by dividing its total clicks by its impressions. Then, for each headline test, we took the mean CTR across variants and divided each variant's CTR by that average. The resulting metric represents a variant's lift relative to the test average. Normalization is necessary because headline tests occurred at different times, in different places on the homepage, and across different domains, making direct comparisons of raw CTRs across tests uninformative.

To illustrate, consider a headline test with three variants whose raw CTRs are 0.02, 0.06, and 0.04 clicks per impression. To calculate lifts, we divide each variant CTR by the test's mean CTR value (0.04), giving lifts of 0.5, 1.5, and 1.0, respectively.

Descriptive Analysis

The distribution of test counts across domains is heavily skewed. Most domains conduct a small number of tests (median = 19), while a couple outliers conduct tens of thousands. This disparity results from variance in content volume and resources available for testing (Hagar and Diakopoulos 2019), as well as differences in when domains first began testing.

Table 1. Count and percentage of tests by number of variants tested. 2-variant tests are most common, and almost all tests have fewer than six variants.

Number of variants	Test count (% of total)
2	103,659 (73.6%)
3	25,578 (18.2%)
4	8,714 (6.2%)
5	2,122 (1.5%)
6+	845 (0.6%)

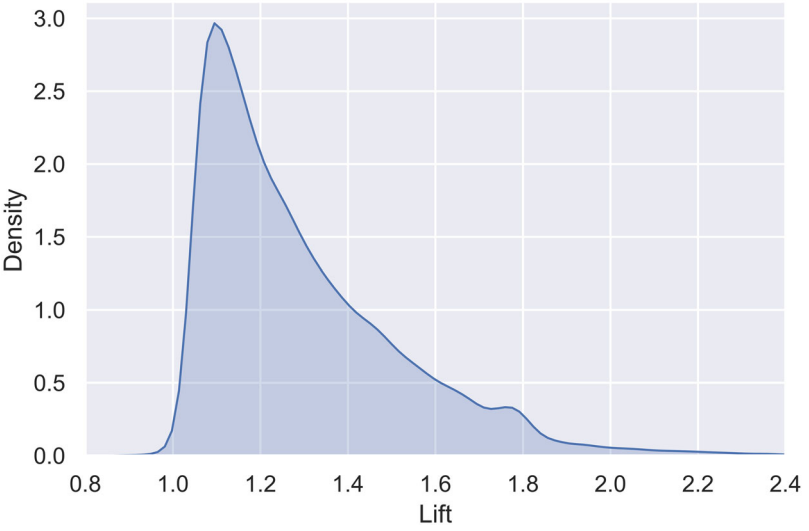

Figure 1. Distribution of lift for winning headlines indicating skew with concentration just above 1, a median lift of 1.23, and a long tail.

Table 1 shows the distribution of the number of headline variants considered in each test. Most tests contain two variants, and very few have more than six.

Figure 1 shows the distribution of lift for winning variants. The median lift for winners was 1.23, indicating that the median winning variant garnered a 23% higher CTR in comparison to the average CTR of the test. Some variants perform far better, with 0.5% of winning headlines showing a lift greater than 2.

Methods

Our analysis uses predictive modeling to understand how headline writing impacts performance. Our processing pipeline operationalizes key features, trains models, and interprets their predictions. To better contextualize our model's performance, we also empirically estimated the upper limit of predictability within our sample.

Feature Engineering

To capture nuanced aspects of headline linguistics and semantics, we leveraged textual features motivated by prior research and theory across four categories: linguistic construction, news values, individual tokens, and semantic embeddings. We also

Table 2. Features and their sources, categories, and number of dimensions.

Feature	Source	Category	# Dimensions
Part of speech	spaCy part of speech labels	Linguistics	50
Named entity type	spaCy entity labels	Linguistics	18
Word count	spaCy token count	Linguistics	1
Character count	String length	Linguistics	1
Mean word length	Character count / token count	Linguistics	1
Fraction of stop words	spaCy stop word labels	Linguistics	1
Reading level	Flesch reading ease score	Linguistics	1
Contains question mark	Pattern matching	Linguistics	1
Contains name	spaCy entity labels	News values	1
Contains A/V term	Pattern matching: custom A/V dictionary and LIWC perception dictionary	News values	1
Contains location term	spaCy entity labels	News values	1
Contains magnitude term	spaCy entity labels	News values	1
Contains shareability term	Pattern matching: custom shareability dictionary	News values	1
Surprise	Empath dictionary cosine similarity	News values	1
Conflict	Empath dictionary cosine similarity	News values	1
Sentiment	VADER dictionary	News values	1
Lemmas	spaCy tokenization	Tokens	825
Semantic embeddings	spaCy large model pre-trained embeddings	Semantic embeddings	300
Datetime	Test metadata	Context	5
Domain	Test metadata	Context	1

incorporated contextual features about each headline test. Table 2 contains the full list of individual features and the source of each computed operationalization.

Linguistics

Linguistic features convey the syntactical components of headlines, including parts of speech and named entities. We also calculated the fraction of a headline made up by stop words (e.g., "the", "of", etc.), whether a headline contains a question mark, and each headline's reading level and length. These features are the basis for several prior empirical studies of headline writing styles (di Buono et al. 2017; Dor 2003; Kuiken et al. 2017).

To identify part of speech types, named entity types (e.g., people and places), and stop words within headlines, we used spaCy, a state-of-the-art software package for Natural Language Processing (NLP) (Honnibal et al. 2020).[4] We indicate whether each part of speech or entity is present in a headline with a binary variable.

To determine a headline's reading level, we calculated its Flesch reading ease score (England, Thomas, and Paterson 1953). Our remaining features—word count, character count, mean word length, and the presence of a question mark—used simple tallies or character searches.

News Values

News values capture theoretical dimensions of how journalists determine and communicate an article's newsworthiness (Harcup and O'Neill 2017). While there is not a universal set of news values across the literature, several qualities (e.g., surprise, conflict, proximity, sentiment) are common (Harcup and O'Neill 2017; Karnowski et al. 2021; Kessler and Engelmann 2019; Parks 2019). Many of these concepts also appear in prior

empirical evaluations of headline performance (Belyaeva et al. 2018; di Buono et al. 2017). Several of our news value operationalizations rely on dictionary approaches. To combat known issues with off-the-shelf language analysis dictionaries (Boukes et al. 2020), we select or create dictionaries specific to particular news values wherever possible.

Studies of news values often involve qualitative and contextual examination of headlines and articles (Harcup and O'Neill 2017). Not all news values transfer well to a quantitative approach because they may require more nuanced examination (e.g., entertainment or topic familiarity—Trilling, Tolochko, and Burscher 2017). Some concern other article elements—such as body text or accompanying visuals—placing them outside the scope of our inquiry. In our analysis, we selected a subset of news values that could be operationalized from headline text: sentiment, reference to the power elite, magnitude, proximity, surprise, conflict, audio/visual signifiers, and shareability.

We calculated headline sentiment using a crowdsourced lexicon of sentiment intensity and valence to score texts with a continuous negative-to-positive measure (Hutto and Gilbert 2014). The approach is designed for short texts and performs as well as or better than comparable lexicons on benchmark evaluations.

To measure references to power elites, we determined the presence of person entities, as labeled by spaCy. Some prior approaches assess the prominence of an identified individual, by measuring traffic to their Wikipedia page, for example (Arapakis, Cambazoglu, and Lalmas 2017). But because our sample contains a diverse range of outlets, we cannot rely on a single measure of prominence to assess a name's newsworthiness. Instead, we assert that any name included in a news headline carries weight for its intended audience based on the journalist's editorial judgement of importance and familiarity of the name to their audience. We also use spaCy to evaluate magnitude, which captures the scope and scale of a story. We identify this from the presence of comparative/superlative adjectives and adverbs, as well as numerical entities (e.g., percentages, ordinal numbers, and counts).

To get at the idea of proximity, we identified headlines with a location from the presence of spaCy's geographic entity labels. Other measures could record the distance between a place and a news organization, to quantify geographic proximity. However, many dimensions of proximal salience (e.g., culture or topical interests) are not captured by distance (Hagar et al. 2020; McCombs and Winter 1981). Even in the strictest geographic sense, physical proximity to an (inter)national news organization tells us little about a location's relevance to news audiences. As such, we eschew geographic proximity in favor of treating any named location as salient to a headline's intended audience. The presence of a location in a headline allows readers to make their own assessment of proximity, which may influence behaviors in a way that leads to patterns we can infer from the data.

Surprise and conflict were both calculated from dictionary expansion, a widely used approach to making lexicons more comprehensive (Gentile et al. 2019). We started with a list of synonyms for "surprise" and "conflict", drawn from Merriam-Webster. We then used Empath, a neural network-based lexical tool, to identify larger groups of related words based on these synonyms (Fast, Chen, and Bernstein 2016). Finally, we calculated headline-level scores for both surprise and conflict based on these

expanded dictionaries. To do so, we relied on *semantic embeddings*, numerical vector representations of words described in more detail below. We use the embeddings from spaCy for each token in the headline and in the dictionary. We then measured the pairwise cosine similarity between every headline token embedding and dictionary token embedding, taking the maximum value of those similarities as the score. This value conveys the extent to which a headline aligns with terms that express surprise or conflict, and it helps to mitigate sparsity issues that might arise from attempting to directly match words in the dictionaries.

Audio/visual signifiers were drawn from the "perception" (273 words) and "see" (72 words) LIWC dictionary categories (Tausczik and Pennebaker 2010). We also augmented these categories with a manually curated list of A/V terms. For shareability, we created a binary indicator for whether the headline contained any matches to a series of phrases identified by industry research on social media shareability (Rayson 2017). Full word lists for our surprise, conflict, A/V, and shareability dictionaries can be found in the Supplemental Materials.

Tokens

Tokens refer to the individual words that appear within headlines. They allowed us to make finer-grained distinctions among categories—not just *whether* a headline has a name, for example, but *which* name. While this level of detail is often difficult to generalize, past research provides support for examining tokens when predicting headline performance (Kim et al. 2016).

We extracted the set of all lemmas from the headlines in our sample using spaCy. Whereas tokens may differ in conjugation or declension (e.g., "run" versus "running"), lemmatization normalizes tokens to their root form. We selected lemmas that were used 100 times or more, and that had a significant ($p < 0.05$) Pearson correlation to lift. For each of the remaining 825 lemmas, we created a binary variable indicating whether it appeared in each headline.

Semantic Embeddings

Word embeddings encode and make comparable the semantics of a text by representing word contexts as dense numerical vectors (Lau and Baldwin 2016). As the product of deep learning models, the individual dimensions of these embedding vectors do not carry inherent conceptual meaning (Shin, Madotto, and Fung 2018). However, word embeddings have proven valuable in headline performance prediction (Lamprinidis, Hardt, and Hovy 2018).

We used the built in pre-trained semantic embeddings from spaCy's large English model (version 2.2.5). These embeddings contain 300 dimensions and were trained on English language text from the OntoNotes 5.0 and GloVe Common Crawl corpora.[5] For each headline, we computed the average embedding vector across all tokens.

Table 3. Number of features retained by category, as well as aggregate feature importance, and some of the specific features retained.

Feature type	Number retained (% of that type)	Total permutation importance	Features retained
Linguistic	3 (4%)	0.009	Average word length, Number of characters, Fraction stop words
News values	3 (38%)	0.010	Surprise, Conflict, Sentiment
Tokens	2 (0.2%)	0.011	"Here", "This"
Semantic embeddings	127 (42%)	0.338	(see text)
Context	1 (17%)	0.001	Domain

Context

Because headline tests occur on dynamic homepages across websites, we also represented broader contextual features that may be relevant to headline performance. We included the (ordinal-encoded) domain as well as the date and time (as year, month, day, day of week, and hour) as additional features.

Modeling

To assess our selected features' ability to predict headline performance, we trained a random forest regressor on a random sample of 75% of our filtered sample, then evaluated it on the remaining 25% using scikit-learn (Pedregosa et al. 2011). Random forests allow us to model potentially complex interactions between features, including non-linear relationships (Hastie, Tibshirani, and Friedman 2009). They also effectively incorporate the mix of continuous, binary, and categorical variables that we utilize, without a need for feature normalization. Finally, random forest models are highly interpretable, providing straightforward measures for the relative importance of each feature in prediction outcomes (Breiman 2001). We compared both standard random forests and gradient boosted trees, and found slightly better performance with the former, which we report here.

We trained a regression model to predict lift at the headline level, then transformed those predictions into test-level ranks (i.e., ranking all variants in a test by predicted performance). We then measured how often the model correctly picked the test winner (precision@1). This approach is analogous to widely used learning-to-rank frameworks (Tatar et al. 2014).

To optimize our model, we ran a grid search over relevant parameters. We varied the number of estimators (50, 100, and 200), minimum samples required to split a node (2, 100, 200, and 400), minimum samples required for a leaf (100, 200, and 400), and the maximum features considered by the model when splitting (square root of total features versus log base two of total features). Our optimal model contained 200 estimators, with split and leaf minimums of 100, and considered log base two number of features when splitting. We also ran 3-fold recursive feature elimination with cross-validation to refine the model's features. Recursive feature elimination removes features one at a time, then evaluates the model's performance on a held-out sample without each feature. It then evaluates each feature's utility to the model (See Table 3 for information on which features were retained in the final model).

Feature Interpretation

We calculated permutation importance to rank the relative contribution of each feature to overall model performance. This approach generates comparable importance metrics that do not depend on the scale or variance of features (Breiman 2001). To examine the relationship between features and performance directly, we also calculated the Spearman correlation between each feature and lift.

Estimated Prediction Ceiling

Content-based performance prediction is a known challenge, because of inconsistencies in audience attention (Arapakis, Cambazoglu, and Lalmas 2017). This inconsistency is driven by factors that cannot be observed through content alone, such as social influence within groups or selective attention heuristics (Salganik, Dodds, and Watts 2006; Zillmann et al. 2004). Because of these difficulties, it is unreasonable to assume that our model's performance would reflect a hard upper bound on predictability. To estimate a more reasonable prediction ceiling, we instead examined the consistency of 4,698 repeated tests within our sample. In these cases, multiple tests were run with the exact same set of headline variants, and at least one test reached hard convergence. Repeat tests may occur because of a user's desire to validate test results. In some cases, tests occurred across multiple domains within the same organization (e.g., networks of local news sites). Because these tests varied along unobserved dimensions, we used them to gauge the impact of outside factors on our model's performance. We define the *replication rate* as the number of repeat test cases that always result in the same winner divided by the total number of repeat tests. The replication rate acts as a rough upper bound for the predictability of headline tests given variance in non-content-related factors. This straightforward calculation is not dependent on our feature engineering or modeling steps, allowing us to establish an estimated prediction ceiling independent of our model's performance.

Results

Content-Based Predictability

We addressed RQ1 (To what extent can a headline's written content predict its performance?) in two parts. First, we evaluated our model's ability to predict headline test outcomes. We measured our model performance using the precision@1 score, which indicates how often the model correctly identifies a test's winning variant. We focus on tests with 2–4 variants in our reporting, since they make up 98% of our sample. Our model's overall precision@1 score was 0.566. That score ranges from 0.4 (for 4-variant tests) to 0.47 (for 3-variant tests) to 0.61 (for 2-variant tests). In all cases, our model outperformed the test-level random baseline—calculated as one divided by the number of variants in a test—by at least 0.11. This performance suggests that it's possible to get some predictive power out of content-based features in this context. Headline writing style at some level does matter and can make the difference between a winning and losing headline. But predictability based on content alone also has clear limits.

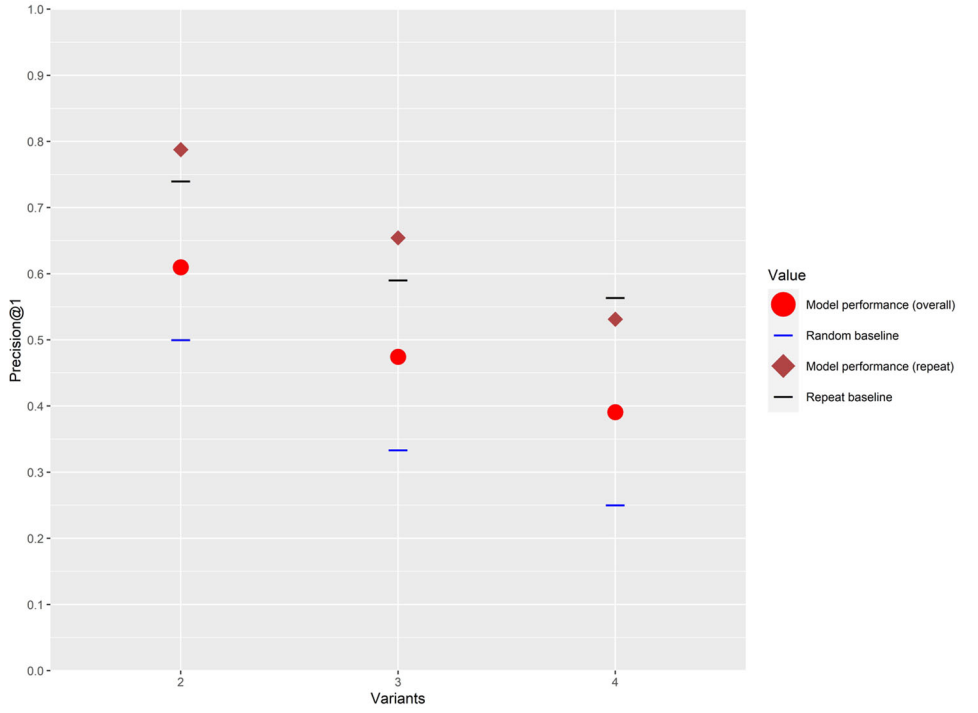

Figure 2. Model performance, compared to random baseline performance and our empirically estimated prediction ceiling. Our overall model performs about halfway between the baselines, while the repeat-only model meets or exceeds the content-only ceiling estimate.

We then examined these limits more closely by calculating a rough upper bound to predictability based on content alone. As described in our Methods section ("Estimated Prediction Ceiling"), we observed the outcomes of repeated tests and computed their replication rate. Values range from 56.4% (for 4-variant tests) to 59.0% (for 3-variant tests) to 74.0% (for 2-variant tests). Figure 2 shows these replication rates (black dashes) against our model's performance. To further validate these estimates, we calculated our model's precision@1 score for only repeat tests. Our model performed within or slightly above the confidence intervals suggested by these replication measurements. The model can exceed this content ceiling because of the contextual information provided by the domain feature, allowing it to account for site-level differences.

These results help establish an estimate of the predictive power of a headline's textual features. While our model achieved modest performance, A/B test outcomes are clearly influenced by factors outside of how a headline is written. Even tests with identical variants change winners anywhere from 26% to 43.6% of the time, pointing to the important role of audience behaviors and preferences as well as other contextual factors in determining test outcomes.

The Impact of Textual Features

We next examined this predictive model in more detail, scrutinizing the impact of individual textual features in response to RQ2 (What is the relative importance of various

Table 4. Feature (Spearman) correlations with lift, as well as permutation importances. All correlations are statistically significant ($^*p < 0.01$), except for the conflict score.

Feature	Spearman correlation	Permutation importance
Average word length	−0.026*	0.004
Number of characters	−0.020*	0.003
Fraction of stop words	0.034*	0.003
Surprise	0.042*	0.005
Conflict	−0.003	0.004
Sentiment	−0.028*	0.001
"Here" (lemma)	0.045*	0.004
"This" (lemma)	0.052*	0.006
Domain	–	0.001

content-based features to headline performance?). We found only marginal importance for any content features. Out of 1,212 features, recursive feature elimination retained 136 (11.2%). By the nature of our chosen model, these results reflect not just the effect of features in isolation, but also their impact relative to every other feature. Valuable features in this context provide information not already captured by other features, allowing us to compare utility to the model across feature categories (Table 3). The linguistic and lemma features fail to provide the model with much useful information beyond what other features offer. In aggregate, the linguistic, news value, and lemma features had comparable value to the model. By far, the most informative category seems to be semantic embeddings. Table 4 contains the Spearman correlation between each (non-embedding) feature and lift. All correlations are statistically significant ($p < 0.01$) except for that of the conflict score ($p = 0.13$). To further unpack the top-level statistics, we next detail the implications of the features chosen within each of our four categories.

Linguistics

Three linguistics features appear in the final model. Contrary to past work, we find that parts of speech and named entities do not impact headline performance in a way that our model can distinguish (Kim et al. 2016; Kuiken et al. 2017). The remaining features' correlations suggest that simpler, less information-dense headlines perform better. Headlines with more stop words correlate with higher performance. As explored in Blom and Hansen (2015), clickbait acts as a forward reference. More stop words mean less substantive information, suggesting that the key content of the article may not be reflected in the headline. The importance of short headlines and shorter words also reflects the findings of Dor (2003).

News Values

The news values features selected by the model—surprise, conflict, and sentiment—focus on headlines' affect. We find a positive association between negative headlines and better performance. These results are in line with evidence of a tendency for people to react more strongly to negative than positive news (Soroka, Fournier, and Nir 2019). While other studies have found that news shareworthiness is predicted more by headline positivity (Berger and Milkman 2012; Trilling, Tolochko, and Burscher 2017),

users may have a different motivation for reading than for sharing an article. Conflict scores have a non-significant negative correlation with performance, while surprise scores have a more substantial positive correlation. The lack of a performance boost from conflict-heavy headlines moderately aligns with past work. Trilling, Tolochko, and Burscher (2017) find a statistically significant but minor increase in sharing for conflict-heavy headlines, while Valenzuela, Piña, and Ramírez (2017) find that a conflict framing reduces the probability that an article will be shared. The news value of "surprise" largely relates to an article's unexpectedness or contrast (Harcup and O'Neill 2017; Kessler and Engelmann 2019). These aspects of surprise are valuable in predicting test outcomes and are correlated with increased headline performance.

Lemmas

Only two words provide the model with meaningful predictive information: "here" and "this". Both lemmas are integral to common headline formulations (e.g., "here's why", "this is how") that are often identified as clickbait (Rayson 2017). They also fulfill the forward-referencing role of clickbait by guiding the reader to a promised piece of information contained either later in the headline or in the text of the article (Blom and Hansen 2015).

Context

As explored in Hagar and Diakopoulos (2019), organizations' testing strategies depend on their priorities, trust in the results of tests, and technical aptitude. Our domain variable implicitly captures these organizational differences. It also acts as a proxy for other variables, such as the size of the outlet, if the domain predominantly covers a particular topic, and facets of behavior that may be specific to its audience. While the domain feature is less important than most others (permutation importance $= 0.001$), its variation adds some site-level nuance to the model's predictions.

Semantic Embeddings

Facilitating clear interpretation and additional theorizing with semantic embeddings is an active area of research. Some extant approaches work to help understand entire embeddings in a relative sense (Kenter and de Rijke 2015; Liu, White, and Dumais 2010), or transform pre-trained embeddings (Panigrahi, Simhadri, and Bhattacharyya 2019). However, we are unaware of methods that elucidate the semantics captured by individual dimensions of pre-trained embedding vectors.

Instead of detailing the individual dimensions of our embedding features and their potential interpretations, we emphasize the utility that the embeddings as a whole provide in this prediction task. They comprise 93% of the features retained by the model (127 of 136), more than any other feature category. They also have the highest permutation importance of any category by far, at 0.338. Their prevalence relative to other features may be a result of the level of granularity they encode. Linguistic categories, for example, tend to encode a relatively broad level of information—labeling a

token as a proper noun provides some information about its contents, but elides many details about linguistic context. Indeed, Dor (2003) makes the distinction between including names and concepts with high news value (which help headline performance) and those that have low news value (which hurt headline performance). In contrast, embeddings can capture nuances related to word context and semantics, which appear to carry the lion's share of value for predicting headline performance.

Discussion

This work explores the intersection of headline writing and audience attention. By examining the predictability of A/B headline tests, we develop ecologically valid insights into how writing strategies impact readers' propensity to click on articles.

We find that the predictability of headline performance based on content alone is limited (RQ1). Even when tests contain identical headlines, their outcomes often vary—up to 26% of the time for two-variant repeat tests. Our model draws some predictive power from textual features, but its performance is still hampered by our inability to measure a headline's context. Headline writing matters, but only to some extent. As noted in past work, content-based prediction proves challenging (Arapakis, Cambazoglu, and Lalmas 2017). There is a rich area of literature demonstrating the non-content factors that impact news reader attention, encompassing theoretical frameworks such as selective exposure, partisan preferences, and social influence (Fischer et al. 2005; Iyengar and Hahn 2009; Lerman and Hogg 2010; Messing and Westwood 2014; Zillmann et al. 2004). Journalists' behaviors can also impact this process, by dictating the type, frequency, and quality of tests they run (Hagar and Diakopoulos 2019).

Our results suggest these external factors play into news audience decision making, even at the micro level of evaluating headlines. They also demonstrate the *dynamic* nature of news audience engagement. Just as news stories shift in salience depending on the issues and events that are prominent at a given point in time, the writing strategies journalists employ to attract audience attention must depend on context (Waldherr 2014). Applying static writing strategies to the fluid nature of online publishing (e.g., across time, duration, position) disregards that situational nature. At least some of a reader's decision to click on a news story occurs outside of the moment of exposure to a headline, requiring a contextual understanding of its presentation. To better understand and predict the relationship between news exposure and engagement, we need to incorporate a story's broader circumstances more explicitly into the study of its reception.

Most individual textual features do not substantially impact our model's predictive power (RQ2), though in aggregate the embedding features, which capture high-dimensional linguistic semantics and context, carry the most predictive power. Many features' correlations with headline performance are directionally consistent with past work, such as sentiment (Soroka, Fournier, and Nir 2019), conflict (Trilling, Tolochko, and Burscher 2017), and surprise (Kessler and Engelmann 2019), but they are weak across the board. Our examination of these features stems from work which theorizes that linguistic and semantic formulations directly impact news engagement (e.g., Kim

et al. 2016; Kuiken et al. 2017). However, given their limited impact, our results caution against over-emphasizing written headline composition when considering the complex factors influencing news attention decisions. While general trends might appear across a large sample for certain features, they do not provide hard and fast rules for improving performance in specific cases. Past work has already established that click behavior is a weak proxy for audience interest in a headline. Readers may click for reasons not related to an article's presentation, and not choosing to click does not equate to a lack of interest (Kormelink and Meijer 2018). Our results reinforce this idea in an A/B testing framework, suggesting a need for more nuanced approaches to testing.

Many newsrooms treat testing as an objective source of optimization data, with a clear relationship between the writing strategies they test and readers' click behavior (Hagar and Diakopoulos 2019). This perception creates an opportunity for ineffective headline writing approaches to gain prevalence, shaping news presentation through a misinterpretation of audience preferences. Similar to the process of writing to an imagined audience, journalists craft their headlines based on an imperfect approximation of reader preferences when relying on A/B test results (Coddington, Lewis, and Belair-Gagnon 2021). Test results (and behavioral analytics more broadly) record data at scale but fail to capture dimensions of audience engagement that are not easily quantifiable (Steensen, Ferrer-Conill, and Peters 2020). These divergent shortcomings complicate our picture of journalistic decision making. Prior work highlights the tension between journalists' professional priorities and the demands of audience metrics (Anderson 2011). Using the framework of the imagined audience, future research should more broadly consider the constellation of audience feedback, as well as the potential shortcomings of its collection or presentation, when evaluating journalists' decisions.

For practitioners, our results stress the importance of adopting A/B testing in newsrooms. In our sample, the median test produces a 23% lift over the average variant click-through rate (even when incorporating less clear-cut soft-converged tests, the median test still generates a 19% lift). Without generalizable best practices for headline writing, continuous testing is the most effective way to achieve this lift because it optimizes the text in relation to specific (potentially unknown) audience and contextual factors. However, newsrooms and testing tool providers must also better communicate the statistical uncertainty of test outcomes and emphasize their context specificity when considering what generalizable lessons can be gleaned by running the tests.

In demonstrating the limitations of content-based prediction, this research suggests a few key areas for future work. First, advances in computational linguistics may allow for more sophisticated encoding of news values. Proximity and power elite could be measured using outside resources (e.g., Arapakis, Cambazoglu, and Lalmas 2017) or models trained via crowdsourced data. As additional signals of audiovisual elements, the images accompanying stories and their contents could be incorporated. Finally, analogous to the shareability measure used in Szymanski, Orellana-Rodriguez, and Keane (2016), fine-tuned language models could provide more sensitive substitutes for the dictionary approaches used to measure conflict and surprise.

Second, more advanced encoding and modeling approaches could improve predictive performance. Given the relative importance of semantic embeddings to our model,

more sophisticated embedding approaches (e.g., sentence embeddings generated by a state-of-the-art model like BERT—Reimers and Gurevych 2019) might provide valuable information about a headline's composition. As new methods arise to interpret these semantic embeddings, we may be able to extract more actionable recommendations for practitioners from them (Panigrahi, Simhadri, and Bhattacharyya 2019). Given neural networks' dominance in natural language processing tasks, they may prove more effective in predicting headline performance (Conneau et al. 2017).

A final area for future work is measuring audience characteristics and behaviors (at the individual level or perhaps clustered into groups—see Makhortykh et al. 2020) and studying how those characteristics interact with the outcomes of A/B headline tests. By examining longitudinal preferences of users through their reading histories and typical consumption patterns, future research might evaluate the consistency of their responses over time to textual elements.

Limitations

This work is also subject to several important limitations that may affect the applicability of our results. First, we only use one class of predictive models, on one subset of data. Other modeling approaches, such as alternative families of regression models or neural networks, may offer increased predictive performance. Our modeling task also only considers hard-converged headline tests. Since soft-converged tests convey a noisier performance signal, our model's performance would likely decline in real-world settings.

Second, we only consider one aspect of an article's presentation. Headlines on a news site homepage are a prominent driver of audience attention, but they are far from the only one. An article might have several distinct headlines—on the home page, the article itself, and social media, for example. While some features we identify agree directionally with work on social media headlines, it is possible that the *magnitude* of their effectiveness varies depending on the source of an article's readership. In future work, cross-source comparison could help quantify the extent to which writing strategy effectiveness varies across audiences.

Finally, our feature engineering approaches introduce several limitations. Because of the scale of our sample, we may overlook words or phrases that are effective for particular audiences, but that our model would not register because of their global sparsity. In addition, several of our features are derived from dictionary-based approaches and may therefore suffer from sparsity. So-called "off-the-shelf" dictionary approaches have well-documented limitations (Boukes et al. 2020; Chan et al. 2021). While we attempt to address them with task-specific dictionary selection, data augmentation, and embeddings to reduce sparsity, our A/V operationalization may suffer from the limitations of the LIWC dictionary. Our measurements of power elites and proximity also rely on straightforward identification of named entities, potentially overlooking distinctions within the classes of people and places identified. Finally, because we focus on news values that can be measured quantitatively, we exclude a handful (e.g., entertainment/drama and relevance) that may influence headline performance (Harcup and O'Neill 2017).

Conclusions

This study presents a large-scale, ecologically valid study of A/B headline tests, challenging the link between headline writing and performance. While practitioners benefit from ongoing A/B testing, our results suggest that they will struggle to obtain generalizable best practices from test results. News audiences are dynamic, and capturing their attention requires more than a staid approach to headline writing. Headline testing, and audience metrics more generally, are only one channel of reader feedback, one that needs proper contextualization and caveats. Equating tracking audience behavior with knowing the audience encourages overreliance on incomplete data, driving flawed approaches to news story presentation.

Notes

1. http://support.chartbeat.com/edu/headlinetesting/methodology.html
2. http://support.chartbeat.com/edu/headlinetesting/orientationguide.html
3. Chartbeat's testing system distinguishes between hard convergence–in which the system is 95% confident that one headline is more successful–and soft convergence. In the latter case, the system selects the variant which it is confident no other headline beats by more than 25%. Because of this relaxed criterion for selecting a winner, soft-converged tests convey a less certain and clear-cut signal of performance for predictive modeling and are therefore excluded.
4. https://spacy.io/api/annotation
5. https://github.com/explosion/spacy-models/releases//tag/en_core_web_lg-2.2.5

Acknowledgements

The authors thank Christopher Breaux, Josh Schwartz, and the Charbeat organization, as well as the reviewers on prior versions of this article for their valuable feedback.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Science Foundation Grant, award IIS-1717330.

ORCID

Nick Hagar http://orcid.org/0000-0001-5110-3737
Nicholas Diakopoulos http://orcid.org/0000-0001-5005-6123

References

Anderson, C. 2011. "Between Creative and Quantified Audiences: Web Metrics and Changing Patterns of Newswork in Local US Newsrooms." *Journalism: Theory, Practice & Criticism* 12 (5): 550–566.

Arapakis, I., B. B. Cambazoglu, and M. Lalmas. 2017. "On the Feasibility of Predicting Popular News at Cold Start." *Journal of the Association for Information Science and Technology* 68 (5): 1149–1164.

Belyaeva, Evgenia, Aljaž. Košmerlj, Dunja Mladenić, and Gregor Leban. 2018. "Automatic Estimation of News Values Reflecting Importance and Closeness of News Events." *Informatica* 42 (4): 527–533.

Berger, J., and K. L. Milkman. 2012. "What Makes Online Content Viral?" *Journal of Marketing Research* 49 (2): 192–205.

Blom, J. N., and K. R. Hansen. 2015. "Click Bait: Forward-Reference as Lure in Online News Headlines." *Journal of Pragmatics* 76: 87–100.

Boukes, Mark, Bob van de Velde, Theo Araujo, and Rens Vliegenthart. 2020. "What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement between off-the-Shelf Sentiment Analysis Tools." *Communication Methods and Measures* 14 (2): 83–104.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Chan, C.-H., J. Bajjalieh, L. Auvil, H. Wessler, S. Althaus, K. Welbers, W. van Atteveldt, et al. 2021. "Four Best Practices for Measuring News Sentiment Using 'off-the-Shelf' Dictionaries: A Large-Scale p-Hacking Experiment." *Computational Communication Research* 3 (1): 1–27.

Coddington, M., S. C. Lewis, and V. Belair-Gagnon. 2021. "The Imagined Audience for News: Where Does a Journalist's Perception of the Audience Come from?" *Journalism Studies* 22 (8): 1028–1046.

Conneau, A., H. Schwenk, L. Barrault, and Y. Lecun. 2017. "Very Deep Convolutional Networks for Text Classification." In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, 2017, 1107–1116. Association for Computational Linguistics.

di Buono, M. P., J. Šnajder, B. Dalbelo Basic, G. Glavaš, M. Tutek, and N. Milic-Frayling. 2017. "Predicting News Values from Headline Text and Emotions." In: Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism, Copenhagen, Denmark, 2017, 1–6. Association for Computational Linguistics.

Diakopoulos, N. 2019. *Automating the News: How Algorithms Are Rewriting the Media*. Cambridge, Massachusetts: Harvard University Press.

Dor, D. 2003. "On Newspaper Headlines as Relevance Optimizers." *Journal of Pragmatics* 35 (5): 695–721.

England, G. W., M. Thomas, and D. G. Paterson. 1953. "Reliability of the Original and the Simplified Flesch Reading Ease Formulas." *Journal of Applied Psychology* 37 (2): 111–113.

Fast, E., B. Chen, and M. S. Bernstein. 2016. "Empath: Understanding Topic Signals in Large-Scale Text." In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose California USA, 7 May 2016, 4647–4657. ACM.

Fischer, Peter, Eva Jonas, Dieter Frey, and Stefan Schulz-Hardt. 2005. "Selective Exposure to Information: The Impact of Information Limits." *European Journal of Social Psychology* 35 (4): 469–492.

Flaxman, S., S. Goel, and J. M. Rao. 2016. "Filter Bubbles, Echo Chambers, and Online News Consumption." *Public Opinion Quarterly* 80 (S1): 298–320.

Fürst, S. 2020. "In the Service of Good Journalism and Audience Interests? How Audience Metrics Affect News Quality." *Media and Communication* 8 (3): 270–280.

Garrett, R. K., and N. J. Stroud. 2014. "Partisan Paths to Exposure Diversity: Differences in Pro- and Counterattitudinal News Consumption." *Journal of Communication* 64 (4): 680–701.

Gentile, A. L., D. Gruhl, P. Ristoski, and S. Welch. 2019. "Explore and Exploit. Dictionary Expansion with Human-in-the-Loop." In *The Semantic Web*, edited by P Hitzler, M Fernández, K Janowicz, et al., Cham, 2019, 131–145. Springer International Publishing.

Hagar, N., and N. Diakopoulos. 2019. "Optimizing Content with a/B Headline Testing: Changing Newsroom Practices." *Media and Communication* 7 (1): 117–127.

Hagar, N., J. Bandy, D. Trielli, Y. Wang, and N. Diakopoulos. 2020. "Defining Local News: A Computational Approach." In Computation + Journalism Symposium, Boston, MA, 2020.

Harcup, T., and D. O'Neill. 2017. "What is News?: News Values Revisited (Again)." *Journalism Studies* 18 (12): 1470–1488.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. "Random Forests." In *The Elements of Statistical Learning*, 587–604. New York, NY: Springer.

Hindman, M. 2015. "Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences." *The ANNALS of the American Academy of Political and Social Science* 659 (1): 48–62.

Honnibal, M., I. Montani, S. Van Landeghem, and A. Boyd. 2020. SpaCy: Industrial-Strength Natural Language Processing in Python. Zenodo

Hutto, C. J., and E. Gilbert. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media.

Ifantidou, E. 2009. "Newspaper Headlines and Relevance: Ad Hoc Concepts in Ad Hoc Contexts." *Journal of Pragmatics* 41 (4): 699–720.

Iyengar, S., and K. S. Hahn. 2009. "Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use." *Journal of Communication* 59 (1): 19–39.

Jungherr, A., O. Posegga, and J. An. 2019. "Discursive Power in Contemporary Media Systems: A Comparative Framework." *The International Journal of Press/Politics* 24 (4): 404–425.

Karnowski, V., D. J. Leiner, A. Sophie Kümpel, and L. Leonhard. 2021. "Worth to Share? How Content Characteristics and Article Competitiveness Influence News Sharing on Social Network Sites." *Journalism & Mass Communication Quarterly* 98 (1): 59–82.

Kenter, T., and M. de Rijke. 2015. "Short Text Similarity with Word Embeddings." In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15, Melbourne, Australia, 2015, 1411–1420. ACM Press.

Kessler, S. H., and I. Engelmann. 2019. "Why Do we Click? Investigating Reasons for User Selection on a News Aggregator Website." *Communications* 44 (2): 225–247.

Kim, J. H., A. Mantrach, A. Jaimes, and A. Oh. 2016. "How to Compete Online for News Audience: Modeling Words That Attract Clicks." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, San Francisco, California, USA, 2016, 1645–1654. ACM Press.

Klinenberg, E. 2005. "Convergence: News Production in a Digital Age." *The ANNALS of the American Academy of Political and Social Science* 597 (1): 48–64.

Kormelink, T. G., and I. C. Meijer. 2018. "What Clicks Actually Mean: Exploring Digital News User Practices." *Journalism (London, England)* 19 (5): 668–683.

Kuiken, J., A. Schuth, M. Spitters, and M. Marx. 2017. "Effective Headlines of Newspaper Articles in a Digital Environment." *Digital Journalism* 5 (10): 1300–1314.

Lamberson, P. J., and S. Soroka. 2018. "A Model of Attentiveness to Outlying News." *Journal of Communication* 68 (5): 942–964.

Lamprinidis, S., D. Hardt, and D. Hovy. 2018. "Predicting News Headline Popularity with Syntactic and Semantic Knowledge Using Multi-Task Learning." In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018, 659–664.

Lau, J. H., and T. Baldwin. 2016. "An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation." In: Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany, 2016, 78–86. Association for Computational Linguistics.

Lerman, K., and T. Hogg. 2010. "Using a Model of Social Dynamics to Predict Popularity of News." In: Proceedings of the 19th International Conference on World Wide Web, Raleigh, North Carolina, USA, 29 April 2010, 621–630. Association for Computing Machinery.

Liu, C., R. W. White, and S. Dumais. 2010. "Understanding Web Browsing Behaviors through Weibull Analysis of Dwell Time." In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2010, 379–386. SIGIR '10. ACM.

Makhortykh, M.,. C. de Vreese, N. Helberger, J. Harambam, and D. Bountouridis. 2020. "We Are What we Click: Understanding Time and Content-Based Habits of Online News Readers." *New Media & Society* 23 (9): 2773–2800.

Margolin, D. B. 2019. "Computational Contributions: A Symbiotic Approach to Integrating Big, Observational Data Studies into the Communication Field." *Communication Methods and Measures* 13 (4): 229–247.

Martin, T., J. M. Hofman, A. Sharma, A. Anderson, and D. J. Watts. 2016. "Exploring Limits to Prediction in Complex Social Systems." In: Proceedings of the 25th International Conference on World Wide Web - WWW '16, Montreal, Quebec, Canada, 2016, 683–694. ACM Press.

McCombs, M. E., and J. P. Winter. 1981. "Defining Local News." *Newspaper Research Journal* 3 (1): 16–21.

Messing, S., and S. J. Westwood. 2014. "Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online." *Communication Research* 41 (8): 1042–1063.

Panigrahi, A., H. V. Simhadri, and C. Bhattacharyya. 2019. "Word2Sense: Sparse Interpretable Word Embeddings." In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, 5692–5705. Association for Computational Linguistics.

Parks, P. 2019. "Textbook News Values: Stable Concepts, Changing Choices." *Journalism & Mass Communication Quarterly* 96 (3): 784–810.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Petre, C. 2015. The Traffic Factories: Metrics at Chartbeat, Gawker Media, and The New York Times. 7 May. Tow Center for Digital Journalism. Accessed 10 December 2018. https://www.cjr.org/tow_center_reports/the_traffic_factories_metrics_at_chartbeat_gawker_media_and_the_new_york_times.php/.

Rayson, S. 2017. We Analyzed 100 Million Headlines. Here's What We Learned (New Research). Accessed 15 October 2018. https://buzzsumo.com/blog/most-shared-headlines-study/.

Reimers, N., and I. Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, 3980–3990. Association for Computational Linguistics.

Ross, A. A. 2017. "If Nobody Gives a Shit, is It Really News?": Changing Standards of News Production in a Learning Newsroom." *Digital Journalism* 5 (1): 82–99.

Salganik, M. J., P. S. Dodds, and D. J. Watts. 2006. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science (New York, N.Y.)* 311 (5762): 854–856.

Shin, J., A. Madotto, and P. Fung. 2018. "Interpreting Word Embeddings with Eigenvector Analysis." In Workshop on Interpretability and Robustness in Audio, Speech, and Language (IRASL), 2018. NeurIPS IRASL.

Shmueli, G. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310.

Soroka, S.,. P. Fournier, and L. Nir. 2019. "Cross-National Evidence of a Negativity Bias in Psychophysiological Reactions to News." *Proceedings of the National Academy of Sciences of the United States of America* 116 (38): 18888–18892.

Steensen, S.,. R. Ferrer-Conill, and C. Peters. 2020. "(against a) Theory of Audience Engagement with News." *Journalism Studies* 21 (12): 1662–1680.

Szabo, G., and B. A. Huberman. 2010. "Predicting the Popularity of Online Content." *Communications of the ACM* 53 (8): 80–88.

Szymanski, T., C. Orellana-Rodriguez, and M. T. Keane. 2016. "Helping News Editors Write Better Headlines: A Recommender to Improve the Keyword Contents & Shareability of News Headlines." Natural Language Processing Meets Journalism. IJCAI.

Tatar, Alexandru, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida. 2014. "From Popularity Prediction to Ranking Online News." *Social Network Analysis and Mining* 4 (1): 174.

Tausczik, Y. R., and J. W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29 (1): 24–54.

Tetlock, P., and D. Gardner. 2015. *Superforecasting: The Art and Science of Prediction*. New York, NY: Crown Publishers.

Trilling, D., P. Tolochko, and B. Burscher. 2017. "From Newsworthiness to Shareworthiness: How to Predict News Sharing Based on Article Characteristics." *Journalism & Mass Communication Quarterly* 94 (1): 38–60.

Valenzuela, S., M. Piña, and J. Ramírez. 2017. "Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing: Framing Effects on News Sharing." *Journal of Communication* 67 (5): 803–826.

Waldherr, A. 2014. "Emergence of News Waves: A Social Simulation Approach: Emergence of News Waves." *Journal of Communication* 64 (5): 852–873.

Webster, J. G. 2016. "The Marketplace of Attention." In *The Marketplace of Attention: How Audiences Take Shape in a Digital Age*. Cambridge: The MIT Press.

Zillmann, D., L. Chen, S. Knobloch, and C. Callison. 2004. "Effects of Lead Framing on Selective Exposure to Internet News Reports." *Communication Research* 31 (1): 58–81.