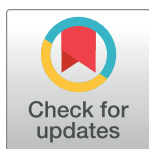RESEARCH ARTICLE

Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report)

Kristina Gligorić[1]*, **George Lifchits**[2], **Robert West**[1], **Ashton Anderson**[2]

1 School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, **2** Department of Computer Science, University of Toronto, Toronto, Canada

* kristina.gligoric@epfl.ch

Abstract

What makes written text appealing? In this registered report, we study the linguistic characteristics of news headline success using a large-scale dataset of field experiments (A/B tests) conducted on the popular website Upworthy.com comparing multiple headline variants for the same news articles. This unique setup allows us to control for factors that could otherwise have important confounding effects on headline success. Based on the prior literature and an exploratory portion of the data, we formulated hypotheses about the linguistic features associated with statistically superior headlines, previously published as a registered report protocol. Here, we report the findings based on a much larger portion of the data that became available after the publication of our registered report protocol. Our registered findings contribute to resolving competing hypotheses about the linguistic features that affect the success of text and provide avenues for research into the psychological mechanisms that are activated by those features.

Introduction

The spread of news and other important information has changed significantly in the age of online social media. As readers increasingly obtain their news over social media [1,2], publishers must engage their readers with individual articles rather than complete newspapers, in what has been dubbed the "unbundling of journalism" [3,4]. Since the same phenomenon gives readers the freedom to obtain news from many sources, publishers are engaged in fierce competition for their readers' attention [4]. Moreover, the nature of online distribution has allowed news organizations to measure engagement at an unprecedented level of granularity and to experiment with distribution methods at a low cost [4–7]. For news publishers, these technological changes have emphasized the importance of crafting an engaging first impression, and have provided the technical infrastructure to conduct rigorous optimization tools for doing so. Publishers, however, ultimately have a limited ability to guarantee the success of their own output and must focus on ensuring that their content is of high quality. For news headlines, this implies developing knowledge of the linguistic predictors of textual success.

There has been substantial scrutiny of the predictors of success in various domains of text production. On our present focus of news, Berger & Milkman [8] studied the characteristics of *New York Times* articles that were heavily shared, identifying that articles that express positive or high-arousal emotions have a higher likelihood of becoming popular. A broad literature focuses on predicting success in news by various means [9–15], although much of this literature prioritizes prediction accuracy above the interpretation of features. The linguistic predictors of success have, however, been studied in other domains. For example, in online social media, Tan et al. [16] discovered several linguistic characteristics of tweets that outperformed closely matched alternatives in an observational study. Other studies on Twitter have investigated how sentiment [17], emotion [18], and length [19,20] affect tweet success. Aside from social media, other studies have used linguistic features to predict success in online communities [21], scientific abstracts [22], literature [23], and quotes [24].

Despite this existing literature, the relationship between linguistic traits and success remains unclear due to fundamental limitations. Broadly, prior work on success employs observational data, where the success outcome can be deeply confounded. Omitted-variable bias can drastically affect the modeled relationship between linguistic covariates and the success outcome to be predicted. For a domain such as news, a number of factors that are often correlated with success can be difficult to control for in observational studies. The time at which content is published can affect success due to concurrent events that create a demand for news or changes in audience size, so any comparison between items that occur at materially different points in time is generally invalid. The author of the content affects success both as a correlate of quality and as a source of social influence. Author skill (though difficult to observe) ought to affect quality, which itself brings success. More importantly, the "superstar phenomenon" [25] demonstrates that the audience for different authors can vary by orders of magnitude, while social influence can affect how an author's content is received, independent of its quality [26]. In a study of Twitter popularity, it was shown that a model including only properties of the tweet author accounted for about half of the optimal model's predictive performance, while the other half was accounted for by the user's past success [27]. Moreover, the content of the article is dependent on the topic it discusses, and different topics have differing audience sizes. The format in which the article is presented also affects its success. In online news, articles on a homepage are typically presented in a grid with a thumbnail image associated with the article. The appeal of the image may drive clicks more than the linguistic properties of the headline. The digital era has enabled some researchers to mine big data for natural experiments that convincingly account for some of these important confounds [16]. It is, however, difficult to fully control for these critically important confounds, which can fundamentally alter the conclusions of any observational study regarding the content-specific predictors of success.

In this Registered Report, a follow-up to a previously-published Protocol [28], we conduct an analysis of experimental field data that provides very strict controls as well as a number of other benefits. We focus on news headlines, which we argue can serve as a "model organism" in an endeavor to elicit the linguistic factors of textual success, as news headlines are specifically crafted to engage with readers at a psychological level. We study a large number of experiments that were conducted by Upworthy, a popular online news publisher, which provides a large sample to test tightly controlled covariates. Each data point of our analysis is a randomized controlled experiment, so all exogenous factors that affect success are strictly controlled for. The experiments cover a long span of time, such that any linguistic covariates of time will be averaged within the multi-year period. Since each experiment varies headline options for a fixed article, and contextual factors such as the thumbnail of the article and the rest of the homepage are also fixed, we can control for the endogenous confounds of author, content, and

context. Finally, the scale of the website on which the experiments were conducted ensures that each experiment is conducted on a large sample and linguistic comparisons are performed with strict measures of statistical significance.

Our analysis is made possible by the Upworthy Research Archive [29], a large dataset of online headline variation experiments made available for research purposes. This rich field experiment data is made available through a partnership between academic researchers and former Upworthy staff. Upworthy was a highly influential online publisher in the U.S. media landscape between 2013 and 2015; in November 2013, Upworthy attracted 80 million unique viewers [30] and was referred to as "the fastest growing media company in the world" [31,32]. With the help of rigorous online experimentation, Upworthy and its contemporaries identified a linguistic style which has since been labeled "clickbait", a recipe so successful at attracting online attention that in November 2016, Facebook publicly announced a modification of their content recommendation algorithms to curb the spread of clickbait [3,33].

The Upworthy Research Archive contains a total of 32,487 headline variation experiments conducted between January 2013 and April 2015 [34]. Each experiment is a comparison of several candidate headline variations authored for a target article, as illustrated in Fig 1. Visitors to the homepage of the Upworthy.com website were shown a selection of articles to view, and for the article that was the subject of any particular experiment, visitors were randomly assigned to see one variant of the headline. Every time the headline variant was shown to a visitor, as well as each time a visitor clicked on that headline variant, the event was logged. This design is referred to as an A/B test [6,7], and its randomized controlled nature allows the experimenter to identify which of several variants has a superior causal effect on clickthrough to its alternatives.

The A/B tests were conducted such that, when the Upworthy homepage was loaded, one article showcased on the homepage was selected for an experiment, with its headline and image varied across experimental conditions. According to former Upworthy engineers, in each experiment only one article on the homepage was varied [33]. Image contents are unavailable in the experimental data, but a unique image ID used for each variation is available, allowing researchers to ensure that the image is held fixed in headline comparisons.

Aside from the unprecedented scale and nature of the dataset, the two-stage process by which this data was made public also follows a novel paradigm. Out of 32,487 total experiments, a time-stratified subset of 4,873 experiments was made available for pilot research. We used this subset as pilot data in order to develop our analysis methodology, form hypotheses, posit the direction and size of the effects, and write the registered report protocol. The remainder of the data became available after the registered report protocol was peer-reviewed and accepted. This release process ensures that all proposed hypotheses are rigorously tested on a
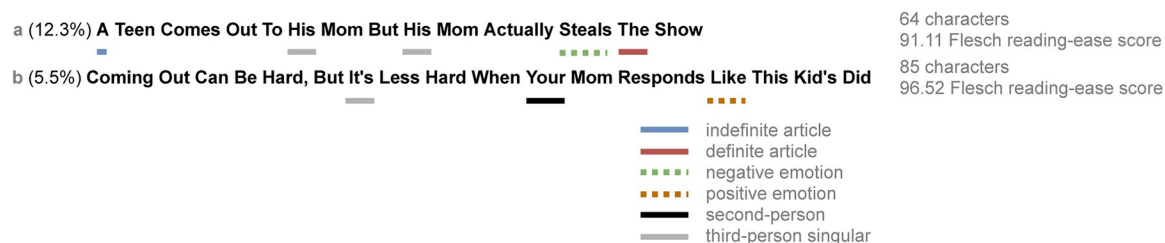
Fig 1. Example of a headline variation experiment pair and derived features. The *a* variant (top) had a higher clickthrough rate (12.3%) than the *b* variant (bottom; 5.5%). The *a* variant contains a definite and an indefinite article, a negative emotion word, and a third-person singular pronoun, whereas the *b* variant contains a positive emotion word, a third-person singular pronoun, and a second-person pronoun. Character count and Flesch reading-ease score are also shown.

https://doi.org/10.1371/journal.pone.0281682.g001

large, unseen dataset without publication bias. The unprecedented scale and experimental nature of the data, together with the scientific rigor of the release process, create an opportunity for conducting valuable confirmatory analyses.

Within the above-described experimental framework, we test hypotheses that we developed based on an exploratory analysis of the pilot dataset and that have been proposed as important factors of success by prior literature. The prior literature is not focused strictly on the domain of news headlines. Similarly, the operationalization of engagement might differ and include sharing instead of clicks. Nonetheless, the studied linguistic cues have been observed as important success factors in similar settings.

Our pre-registered analyses assess eight specific hypotheses. The hypotheses, with the respective sampling plan, analysis plan, and the planned interpretation of the outcomes are summarized in the Design Table (Table 1).

Predictability

The first hypothesis validates the basic premise of our analyses: Is it possible to attribute headline success to the linguistic features of headlines? Success can be the consequence of many complex factors at play, many of which are not observable or subject to unpredictable external shocks [35,36]. It has been shown that even a fully-described complex system can be so prone to the accumulated effects of random behavior that reasonable predictability is impossible [26,27,37]. It is therefore not *a priori* clear that the success of content in complex sociotechnical systems can be predicted at all. However, the experimental nature of the Upworthy Research Archive data allows us to precisely control for time and topic, which accounts for complex social factors. Therefore, any differences in headline success should be almost entirely accounted for by the individual decisions of consumers, and by ensuring that paired comparisons have a sufficiently large sample size, the unobservable factors that affect individual-level decisions are averaged out. Our first analysis thus explicitly asks: Is there any systematic variation in the success of headlines that can be explained or predicted based on linguistic features?

H1: The more successful headline in a controlled pair can be predicted at a statistically significant level based on linguistic features.

Following this first high-level hypothesis, we next turn our attention to specific hypotheses about the individual linguistic factors of headline success and discuss literature which supports them.

Emotion

The use of emotional wording has been explored in several contexts. Prior work suggests that the use of emotional words increases sharing probability in several contexts, such as newspaper articles [8] and tweets [16,18]. Furthermore, the type of emotional reaction elicited by a text may affect its success. Past work supports conflicting views about broadly what kind of emotion is inherently more appealing to individuals. For example, the "Pollyanna hypothesis" [38] states that there is a human tendency to use positive language, a result that has been empirically verified across vast and diverse corpora [39]. In the online sphere, positive affect is more prevalent than negative affect on Twitter, indicating that people generally tend to tweet about happy things [40].

On the other hand, a general result suggests that bad events have greater power than good ones over a wide range of psychological scenarios, including that bad events elicit more information processing, stronger memory, and have more pronounced effects on impression

Table 1. Design table.

Question	Hypo-thesis	Sampling Plan	Analysis plan	Interpretation given to different outcomes
The more successful headline in a controlled pair can be predicted at a statistically significant level based on linguistic features.	H1	See †	Evaluate test accuracy on the Confirmatory Dataset. The regression weights will be the same as those obtained in the Pilot data regression	Null hypothesis: classification accuracy is not significantly better than a majority vote classifier. Accept H1 if the Pilot data regression obtains significantly better accuracy on the Confirmatory Dataset headline pairs with α = 0.01.
The presence of positive-emotion words is negatively associated with headline success.	H2a	See ‡	Examine the regression coefficient "positive emotion"	Accept H2a if the coefficient is negative and statistically significant with α = 0.01, reject otherwise.
The presence of negative-emotion words is positively associated with headline success.	H2b	‡	Examine the regression coefficient "negative emotion"	Accept H2b if the coefficient is positive and statistically significant with α = 0.01, reject otherwise.
Length is positively associated with headline success.	H3	‡	Examine the regression coefficient "number of characters"	Accept H3 if the coefficient is positive and statistically significant with α = 0.01, reject otherwise.
Higher readability is negatively associated with headline success.	H4	‡	Examine the regression coefficient "Flesch reading-ease score"	Accept H4 if the coefficient is negative and statistically significant with α = 0.01, reject otherwise.
Generality (the use of indefinite articles) is positively associated with headline success.	H5a	‡	Examine the regression coefficient "indefinite article"	Accept H5a if the coefficient is positive and statistically significant with α = 0.01, reject otherwise.
Specificity (the use of the definite article) is negatively associated with headline success.	H5b	‡	Examine the regression coefficient "definite article"	Accept H5b if the coefficient is negative and statistically significant with α = 0.01, reject otherwise.
The use of first-person singular pronouns (referring to the author) is positively associated with headline success.	H6a	‡	Examine the regression coefficient "first-person singular"	Accept H6a if the coefficient is positive and statistically significant with α = 0.01, reject otherwise.
The use of first-person plural pronouns (referring to the author and the reader) is negatively associated with headline success.	H6b	‡	Examine the regression coefficient "first-person plural"	Accept H6b if the coefficient is negative and statistically significant with α = 0.01, reject otherwise.
The use of second-person pronouns is positively associated with headline success.	H7	‡	Examine the regression coefficient "second-person"	Accept H7 if the coefficient is positive and statistically significant with α = 0.01, reject otherwise.
The use of third-person singular pronouns is positively associated with headline success.	H8a	‡	Examine the regression coefficient "third-person singular"	Accept H8a if the coefficient is positive and statistically significant with α = 0.01, reject otherwise.
The use of third-person plural pronouns is positively associated with headline success.	H8b	‡	Examine the regression coefficient "third-person plural"	Accept H8b if the coefficient is positive and statistically significant with α = 0.01, reject otherwise.

† Performed on all headline pairs in the Confirmatory Dataset obtained following analysis pipeline described in Methods. ‡ All headline pairs obtained following analysis pipeline are scored on the features, and regression analysis are performed on all pairs. Note that accepting a hypothesis H1-H8 entails rejecting the corresponding null hypothesis.

https://doi.org/10.1371/journal.pone.0281682.t001

formation [41]. In the news domain, negative information has been shown to have a greater impact on individuals' attitudes than positive information [42]. Politically interested participants were found more likely to select cynical and negative news frames [43], which elicit stronger and more sustained reactions than positive news [44]. The negativity bias might be further amplified in digital media [45].

For social media, it has been shown that positive content may receive more popularity than negative content [17,21], but negative messages spread faster [17], and negative words are

more likely to be perceived as relevant to success [20]. Based on these latter results and our findings from pilot data, we formed the following two hypotheses:

H2a: The presence of positive-emotion words is negatively associated with headline success.

H2b: The presence of negative-emotion words is positively associated with headline success.

Length

The appeal of a headline may be associated with its length via competing factors. A Gricean maxim of cooperative communication emphasizes that the *quantity* of transmitted information should be sufficient to be informative, but only as much as required [46]. Meanwhile, the maxim of *relation* may favor brevity, introducing a tension between being informative and being concise [47,48]. There are reasons why shorter headlines may be expected to perform better. Shorter posts were found to be more successful on Twitter, with length constraints improving tweet quality [19], and by shortening original tweets to various lengths, Gligorić et al. [20] found that tweets which are up to 30–40% shorter than their longer original versions are more likely to be judged as successful. Accordingly, in a study of phrases of text being repeated in various online sources, Simmons et al. [49] found that shorter phrases were used more often. On the other hand, a longer headline may contain more information, with a higher probability of engaging the reader [16], a hypothesis that was indeed supported by our analysis of pilot data:

H3: Length is positively associated with headline success.

Readability

Highly readable text may be more sympathetic to the reader, while less readable text may provide more information. A matched observational study of topic- and author-controlled tweets revealed that tweets with higher readability are more likely to be successful [16]. However, the linguistic style of text posted on Twitter is substantially different from text present in other corpora of other online content such as online blogs [50] or the news headlines studied here. In particular, Hu et al. [40] described how stylistic features correlated with readability vary significantly across media. In a study of successful literary works, Ashok et al. [23] found that readability was negatively associated with success. A proposed explanation is that great literature demonstrates high conceptual complexity, which in turn demands lower readability. For the present domain of news headlines, our preliminary analysis of pilot data yielded no significant effect. Aligned with the work of Ashok et al. we therefore state the following hypothesis:

H4: Higher readability is negatively associated with headline success.

We measure readability with the Flesch reading ease score [51], which decreases as either words per sentence increase or syllables per word increase. Thus, higher values imply that the text is more readable.

Generality

Broadly speaking, the use of indefinite articles ("a", "an") can signal generality in the subject discussed [24], whereas the definite article ("the") typically makes reference to something specific and unique [52]. Regarding the appeal of text, Danescu-Niculescu-Mizil et al. [24] found that the usage of general language made movie quotes more likely to be remembered.

Similarly, Tan et al. [16] found in their study of topic- and author-controlled tweets that the inclusion of indefinite articles had a positive effect on tweet success. Our pilot analysis yielded no evidence that the usage of these articles has a significant effect on headline success. Based on the existing work, we thus formed the following hypotheses:

> *H5a: Generality (the use of indefinite articles) is positively associated with headline success.*

> *H5b: Specificity (the use of the definite article) is negatively associated with headline success.*

Pronouns

The use of pronouns in a headline can indicate whether the headline is inclusive of the reader, the author, or refers to a third party. Different pronouns can significantly alter the tone of the headline and certain pronouns may be broadly preferable to readers in general. For instance, Ashok et al. [23] found that pronouns were associated with highly successful books. We consider first-, second-, and third-person pronouns separately. First-person pronouns in particular have been found to contribute to success in scientific abstracts [22], but were not found to correlate with success in tweets [16]. In our pilot analyses, there was a significant positive effect of the inclusion of first-person singular pronouns, which refer to the author, but a negative and non-significant effect of the inclusion of first-person plural pronouns, which refer to a collective that may include both the author and the reader.

> *H6a: The use of first-person singular pronouns is positively associated with headline success.*

> *H6b: The use of first-person plural pronouns is negatively associated with headline success.*

Second-person pronouns refer directly to the reader (i.e., "you"). A prediction study of news headline popularity found that second-person pronouns are associated with more popular headlines [14]. A study of the success of songs found that the use of second-person pronouns is empirically correlated with song success and has a positive causal effect on people liking a song [53]. Our pilot analysis yielded a non-significant positive effect of second-person pronouns on headline success. Aligned with previous work, we formed the following hypothesis:

> *H7: The use of second-person pronouns is positively associated with headline success.*

Third-person pronouns were found to have a positive effect on tweet success [16]; they were, however, not found to be associated with popularity in a study of news headlines [14]. To the best of our knowledge, no prior work has found a significant distinction between third-person singular and plural pronouns. Our analysis of pilot data found that third-person *singular* pronouns (i.e., "she", "his") were positively associated with more engaging headlines, whereas third-person *plural* pronouns (i.e., "they", "theirs") were positively and non-significantly associated with headline success, so we hypothesized:

> *H8a: The use of third-person singular pronouns is positively associated with headline success.*

> *H8b: The use of third-person plural pronouns is positively associated with headline success.*

Methods

The use of exploratory and confirmatory datasets

At a high level, our work involves two analyses that hinge on one logistic regression model: the first analysis aims to determine whether the model has meaningful out-of-sample predictive

accuracy, whereas the second analysis interprets the regression coefficients to assess factors that are associated with headline performance. The release schedule of the Upworthy Research Archive was intended to prevent scientific methodological errors that threaten the validity of hypotheses formed based on the data (such as p-hacking or cherry-picking subsets of the data). In this section we describe how our analysis makes use of each portion of the dataset.

Note that we did not propose any exploratory analyses in the registered report protocol. Our use of the phrase "Exploratory Dataset" follows terminology from the Upworthy Research Archive team, and simply refers to the initial stage of the Upworthy Research Archive data release. We treat the Exploratory Dataset as the pilot data based on which we designed our methodology and formed our hypotheses.

H1: Evaluation of predictability of headline success.   A common issue in statistical learning is *overfitting*, in which an estimator exploits associations between predictors and the outcome that are idiosyncratic to the training data. Since there is a high probability of random associations occurring, an overfitted estimator will have high accuracy within the training sample. However, the goal of most statistical learning applications is *generalization*, or finding rules that yield good predictive performance on unseen data. Techniques such as cross-validation are designed to estimate generalization accuracy [54], but the most reliable assessment uses a large portion of the dataset that was never used in the training process. We therefore evaluate H1 by testing the predictive performance of the logistic regression model trained on the *Exploratory Dataset*, using the Confirmatory Dataset as a large held-out testing dataset.

H2-H8: Evaluation of linguistic hypotheses.   Our second analysis involves the interpretation of logistic regression coefficients to probe the specific meaning of effects that are observed in the data. For this analysis, we fit a logistic regression to the Confirmatory Dataset and analyze the coefficients as described in the Design Table (Table 1).

Design

We study the linguistic traits of headlines by examining how the presence of words increases the odds of a headline being considered better than its alternative. The unique randomized experimental setup in which the data was collected enables this research design by allowing one to disregard any omitted variables that are causally relevant to headline success. According to the Upworthy Research Archive team [33], the original assignment of readers to experimental conditions was random, and only the headline and article image was visible to Upworthy readers as part of any headline variation test. By controlling for headline variations with the same image and conducted within the same week, we ensure that any differences in headline success are fully accounted for by the differences in words used in the headlines themselves.

The Upworthy Research Archive consists of data on online headline variation experiments. Most of these experiments test several headline variations for any given article. Every time a specific headline variant is shown to a reader, it is counted as an *impression*, and when the headline variant is clicked it is counted as a *click*. The *clickthrough rate* for any particular headline variant is defined as *clicks* divided by *impressions*. Experiments can vary other properties, but only the image ID, headline, and week during which the test was conducted is relevant to the impressions received on the Upworthy homepage [33].

Our research hypotheses require data about headlines that are better than a comparable alternative. We obtain pairs of headline variants by considering all possible pairs within each headline variation experiment, such that any headline pair under consideration has the same article ID and image ID, and was tested in the same week. Within each pair of headlines obtained this way, we define as "better" the headline with the higher clickthrough rate, and as "worse", its counterpart.

Sampling plan

Groups of controlled experiments include varying numbers of comparable headlines that can be paired into comparison pairs. Within a group of controlled experiments, we start by considering every possible pair of comparison headlines. In case with more than $K = 15$ headline comparison pairs within an experiment, we randomly sample a subset of $K = 15$ comparison headline pairs. We have run the complete analysis pipeline for different values of K, obtaining the same effect directions and comparable effect sizes. We perform sampling of comparison pairs within an experiment in order not to skew the estimates given idiosyncrasies of particular experiments, since specific experiments relate to articles covering different events and different topics. Note that we do not include a random variable for the experiment since one of the objectives is to apply the model previously fitted on the pilot data directly to the confirmatory data containing unseen experiments.

For many comparison pairs in the data, the better headline performed only marginally better than the worse headline, while for other pairs, there were only few impressions received by one headline variant. Then, within a group of controlled experiments, we perform a Pearson chi-squared test on the clickthrough rates for every possible pair. With each pairwise comparison in a headline experiment there is a probability of incorrectly rejecting the null hypothesis, so we apply the Bonferroni correction [55] with a family-wise error rate of $\alpha = 0.05$. Among such possible matchings of comparable headlines into distinct pairwise comparisons, we select the configuration with the lowest Bonferroni-corrected p-value. After this process, in experiments testing more than two comparable headlines, each headline participates in a single comparison pair.

When testing our hypotheses, the unit of analysis is a pair of comparable headlines, such that each headline among the analyzed set of pairs is unique. All hypotheses presented in this report were developed with the Exploratory Dataset release of the Upworthy Research Archive. With the Exploratory Dataset, we obtained 5,048 pairs of comparable headlines and performed an initial test of all hypotheses on this set of headline pairs. With the much larger Confirmatory Dataset, we obtained 24,333 pairs of comparable headlines and tested the hypotheses on this held-out set. To support time-series research, both Confirmatory and Exploratory Datasets are a random sample of A/B tests, stratified by week number. All hypotheses described in this report are tested on the Confirmatory Dataset using the pre-registered Analysis Plan, but with a large sample of unseen data. All data pre-processing steps are unchanged from what was developed on the Exploratory Dataset and described in the initial protocol.

Analysis plan

Since linguistic features are the primary object of study, our analysis focuses on counting words used in the headline pairs. We developed a dictionary of specific words which we considered for the set of pronoun categories, and used "a" and "an" for the *indefinite article* category (full dictionaries are available in Table 2). For the positive and negative emotion categories, we used Linguistic Inquiry and Word Count (LIWC) [56] to categorize individual words as possessing either positive or negative emotion. Finally, the *textstat* library for Python [57] is used to compute the Flesch reading-ease score [51] and the number of characters for each headline. This process is used to obtain a feature-vector encoding for each headline in the dataset. The process is depicted in Fig 1.

For each pair, we then compute feature vectors for the better and worse headlines. The feature encoding for each headline pair is the difference between the better headline's features and the worse headline's features. All linguistic features merely count the presence or absence of any words in the headlines: thus, for a headline pair, the linguistic feature is 1 if it only

Table 2. Hypothesis word dictionaries.

Category	Words
first-person singular	i, i'd, i'll, i'm, i've, id, im, ive, me, mine, my, myself
first-person plural	our, ours, ourselves, us, we, we'd, we'll, we're, we've
second-person	ya, you, you'd, you'll, you're, you've, youll, your, youre, yours, yourself, yourselves, youve
third-person singular	he, he'd, he's, her, hers, herself, hes, him, himself, his, it, its, itself, she, she'll, she's, shes, themself
third-person plural	their, theirs, them, themselves, they, they'd, they'll, they've, theyll, theyve
indefinite article	a, an
definite article	the

https://doi.org/10.1371/journal.pone.0281682.t002

occurs in the better headline, -1 if it only occurs in the worse headline, and 0 if it occurs in neither or both headlines. The number of characters and the Flesch reading-ease score are real numbers, so the number-of-characters feature is 1 if the better headline is longer than the worse headline, -1 if the worse headline is longer than the better headline, and 0 if the two headlines are equally long; and the reading-ease feature is 1 if the better headline is easier to read than the worse headline, -1 if the worse headline is easier to read than the better headline, and 0 if the two headlines are equally readable. This constitutes the design matrix of predictors.

An outcome vector of length N containing half zeros and half ones is generated and permuted; each row of features is then multiplied by -1 if the outcome is 0 and remains the same if the outcome is 1. This sets up a binary classification problem with perfectly balanced classes, such that either a majority-vote or random classifier would obtain 50% accuracy on this prediction task. For the linear model that we use in our analyses, a positive coefficient for a feature means that the feature was more prevalent in the better headline, whereas a negative coefficient means that the feature was more prevalent in the worse headline.

Each hypothesis in our report is defined by either a set of tokens, a deterministic rule for selecting tokens, or a deterministic function from headline text to output value. Our design matrix thus has one column to quantify each hypothesis. We fit a logistic regression on this design matrix and analyze the coefficients. As described in our Design Table (Table 1), each coefficient maps to a hypothesis. When analyzing the Confirmatory Dataset, we say that a hypothesis is supported if its corresponding column in the regression has $p < 0.01$. For our pilot analyses, we considered $p < 0.05$ as preliminary evidence for the hypotheses.

Results

Following the pre-registered analysis plan, the dataset of confirmatory headline pairs is constructed, features are computed, and a logistic regression model is trained. There are no deviations from the pre-registered analysis plan.

Registered analyses

Our first hypothesis, H1, posits that our model can predict headline success from linguistic features significantly better than random guessing. We evaluate H1 by testing the predictive performance of the logistic regression model trained on the Exploratory Dataset, using the Confirmatory Dataset as a large held-out testing dataset.

Using our regression model, with a 0.5 decision threshold on the predicted value, prediction accuracy on the Confirmatory Dataset is 54.42% (*Pearson* $\chi^2(1)$ = 94.99, $P<10^{-6}$, *odds ratio* = 1.19, 99% *CI* = [0.536, 0.552]), recall is 53.17%, and precision is 54.53%. We thus find

evidence in favour of the hypothesis that the more successful headline in a controlled pair in the Confirmatory Dataset can be predicted significantly better when using linguistic features than when guessing randomly.

To access the remaining hypotheses H2-8, we fit a logistic regression to the Confirmatory Dataset and analyze the coefficients as described in the Design Table (Table 1). The fitted regression coefficients are depicted in Fig 2. Regression coefficients estimated based on the Exploratory Dataset are displayed for reference. Detailed statistics are presented in Table 3. Overall, out of the eleven tested linguistic hypotheses, confirmatory analyses reveal evidence consistent with seven, whereas we fail to find evidence in favour of four, following the pre-registered interpretation outlined in Table 1. Note that here we refer to the consistency between confirmatory analyses and the investigated hypotheses (Table 1), and not to the consistency between confirmatory analyses and the exploratory analyses (Table 4).

Emotion.   Regarding positive-emotion words (H2a), we hypothesized that their presence is negatively associated with headline success (H2a). We find no evidence in favour of this hypothesis ($\beta = -0.04$, $p = 7.15 \times 10^{-2}$) as no significant negative effect is detected. Regarding negative-emotion words (H2b), we hypothesized that their presence is positively associated with headline success. Confirmatory analysis supports this hypothesis ($\beta = 0.18$, $p = 1.61 \times 10^{-14}$).

Length.   Consistent with the hypothesis H3, we find that length is positively associated with headline success ($\beta = 0.07$, $p = 7.57 \times 10^{-8}$).
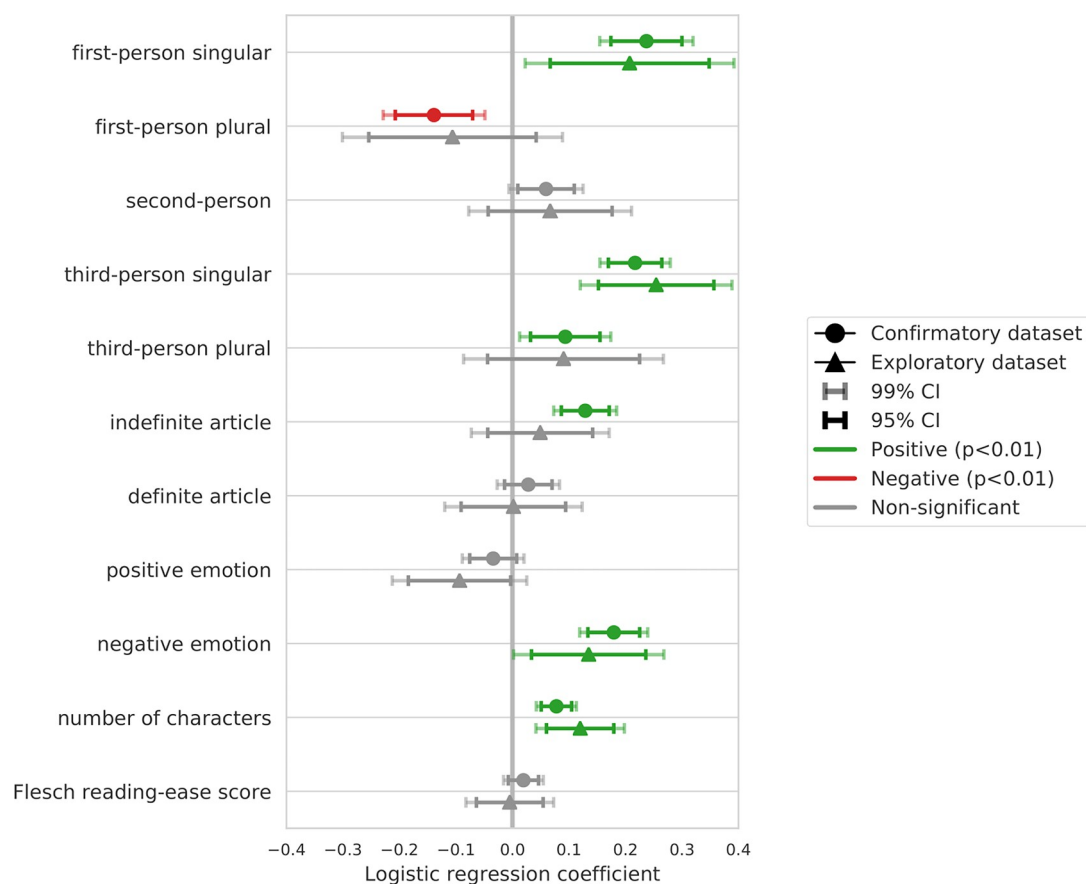
Fig 2. Main regression analysis on the upworthy archive confirmatory and exploratory datasets. Error bars visualize 95% (dark bars) and 99% (light bars) confidence intervals on the logistic regression coefficients.

https://doi.org/10.1371/journal.pone.0281682.g002

Table 3. Main analysis regression with exploratory dataset. Logit regression analysis for confirmatory data.

	Dep. Variable:		y	No. Observations:		24333
	Model:		Logit	Df Residuals:		24332
	Method:		MLE	Df Model:		10
	Date:		Fri, 18 Feb 2022	Pseudo R-squ.:		0.00969
	Time:		16:34:20	Log-Likelihood:		-16703
	Converged:		True	LL-Null:		-16866
	coef	std err	z	P>|z|	[0.025	0.975]
first-person singular	0.2407	0.032	7.5	0	0.178	0.304
first-person plural	-0.1487	0.035	-4.256	0	-0.217	-0.08
second-person	0.0513	0.025	2.017	0.044	0.001	0.101
third-person singular	0.2158	0.024	8.951	0	0.169	0.263
third-person plural	0.0937	0.031	2.984	0.003	0.032	0.155
indefinite article	0.1246	0.022	5.763	0	0.082	0.167
definite article	0.0331	0.021	1.544	0.123	-0.009	0.075
positive emotion	-0.0382	0.021	-1.802	0.071	-0.08	0.003
negative emotion	0.1798	0.023	7.678	0	0.134	0.226
number of characters	0.0739	0.014	5.377	0	0.047	0.101
Flesch reading-ease score	0.0187	0.014	1.369	0.171	-0.008	0.045

https://doi.org/10.1371/journal.pone.0281682.t003

Readability. We hypothesized that higher readability is negatively associated with headline success (H4). We do not find evidence in favour of this hypothesis ($\beta = 0.02$, $p = 1.71 \times 10^{-1}$) as no significant negative effect is detected.

Generality. We hypothesized that generality, i.e., the use of indefinite articles is positively associated with headline success (H5a), while specificity, i.e., the use of definite articles is negatively associated with headline success (H5b). We find evidence in support of H5a ($\beta = 0.12$, $p = 8.26 \times 10^{-9}$), but we do not find evidence in support of H5b ($\beta = 0.03$, $p = 3.31 \times 10^{-2}$).

Pronouns. Out of the five hypotheses related to pronouns, our confirmatory analyses are consistent with four, and not consistent with one. The use of first-person singular pronouns (H6a; $\beta = 0.24$, $p = 6.37 \times 10^{-14}$) is positively associated, whereas the use of first-person plural (H6b; $\beta = -0.15$, $p = 2.08 \times 10^{-5}$) pronouns is negatively associated with headline success. Furthermore, the use of third-person pronouns, both singular (H8a; $\beta = 0.22$, $p = 3.52 \times 10^{-19}$) and plural (H8b; $\beta = 0.09$, $p = 2.85 \times 10^{-3}$) is positively associated with headline success. On the contrary, we found no significant positive association between the use of second-person pronouns and headline success (H7; $\beta = 0.05$, $p = 4.37 \times 10^{-2}$).

Table 4. Outcomes of the tested linguistic hypotheses, depending on the sign of the corresponding coefficient and its p-value. The entries on the diagonal represent aligned interpretations on confirmatory and exploratory datasets (H1, H2a, H2b, H3, H4, H5b, H6a, H7, and H8a). Off-diagonal entries represent discordant results on confirmatory and exploratory datasets (H5a, H6b, and H8b).

		Confirmatory dataset		
		Positive (p<0.01)	Negative (p<0.01)	Non-significant
Exploratory dataset	Positive (p<0.01)	H2b, H3, H6a, H8a	—	—
	Negative (p<0.01)	—	—	—
	Non-significant	H5a, H8b	H6b	H2a, H4, H5b, H7

https://doi.org/10.1371/journal.pone.0281682.t004
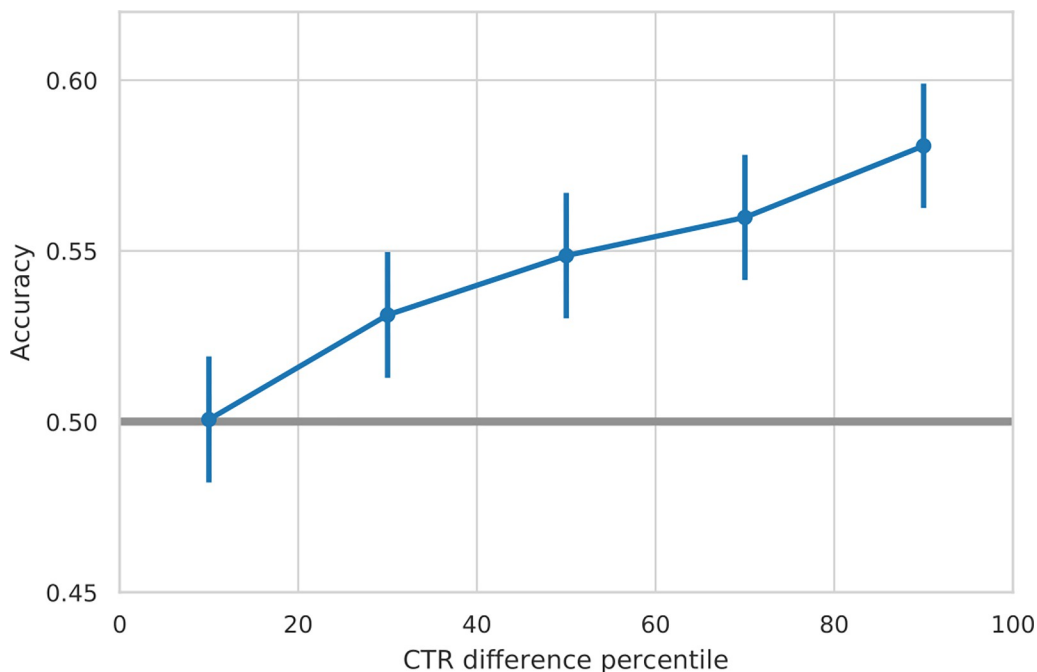
Fig 3. Accuracy vs. CTR. Accuracy (on the y-axis), for ranges of difference in click-through rate between the compared headlines (on the x-axis), split into quintiles.

https://doi.org/10.1371/journal.pone.0281682.g003

Exploratory analyses. In addition to the main confirmatory analysis described above, we perform exploratory post hoc analyses that allow us to understand further the performance of the logistic regression model and the tested headlines.

Dose-response relationship. Given the ease of interpretation, our pre-registered analyses treat success prediction as a binary classification, therefore treating large differences and small differences in the click-through rate (CTR) between the compared headlines the same. We have further studied the performance of the regression model in identifying the better headline by exploring the presence of a dose-response relationship. In our hypothesis H1, we only predict that the model will perform significantly better than the random baseline when making predictions on the complete dataset. However, we also expected that the model would perform much better in instances when the difference in click-through rate (CTR) between the compared headlines is larger. Performing such an analysis (Fig 3), we see that, indeed, the model performs better when the difference in CTR is higher. The peak achieved accuracy is 58.3%.

Headline topics. We investigated the topic distribution of the headlines (Fig 4) by tagging the headlines with the default Empath [58] topics trained on *New York Times* headlines. The most frequent topics among the tested headlines are related to everyday ordinary experiences and entertainment news (e.g., speaking, children, communication, listening, school, or family). The nature of the tested headlines should be taken into account when considering the extent to which our findings are expected to generalize to other news domains (e.g., politics or sport).

Positive and negative emotion. While the specific words associated with the hypotheses related to pronouns and articles are stated in Table 2, we further explored what specific words carry positive or negative emotions. For the positive and negative emotion categories, the top 20 words most frequently categorized as either positive or negative emotion in the confirmatory dataset are listed in Table 5. The presence of the listed words might not necessarily be an accurate proxy for the emotional charge of the headline. For instance, the word "like" can be
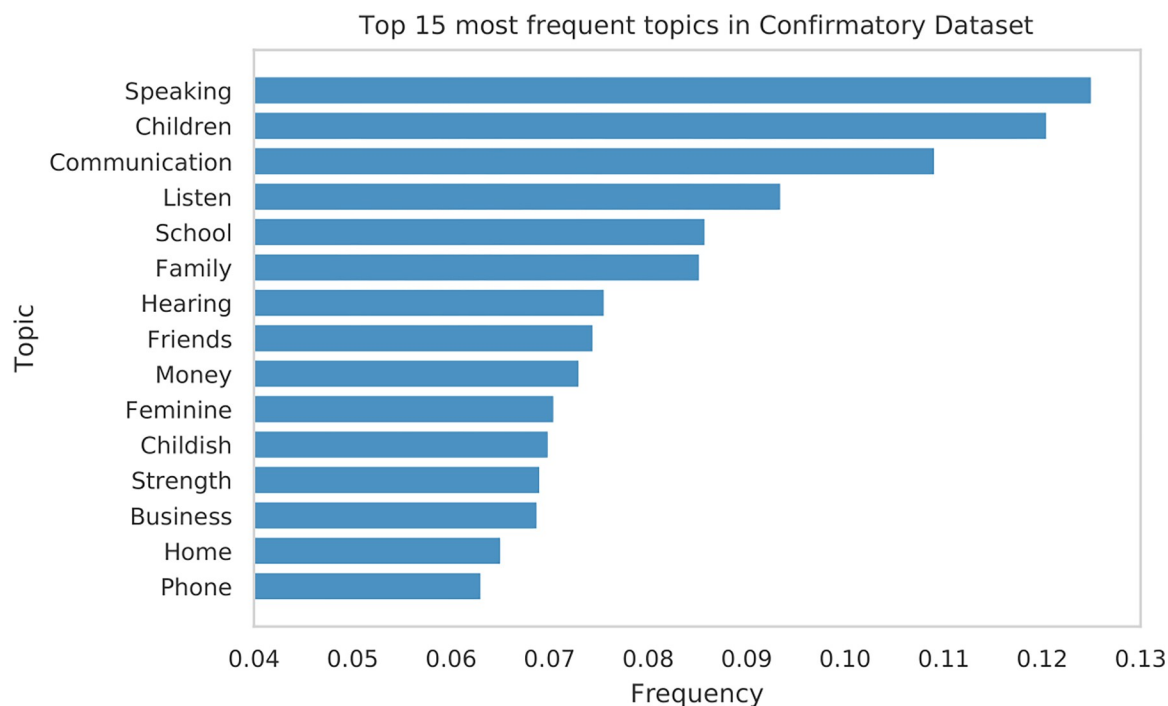
Fig 4. Topics. Top 15 topics by frequency among the confirmatory headlines. The frequency is measured as the fraction of headlines labelled with the respective Empath topic.

used as a linker word, without implying a positive emotion. Therefore, we further explored the impact that the presence of each specific frequent word listed in Table 2 has on the estimated effect in the confirmatory dataset.

For each word frequently categorized as either positive or negative emotion, we individually omitted that word from the category dictionary, fitted the regression model following the analysis plan, and estimated the corresponding coefficients for positive and negative emotion. Regarding positive-emotion words (H1b), in the main confirmatory analyses, we found no evidence in favor of the hypothesized effect (β = -0.04, p = 7.15×10^{-2}). We found that this conclusion is robust to the exclusion of any individual word from the dictionary, with the effect estimate ranging between β = -0.04 (corresponding to the variation when the word "pretty" is excluded) and β = -0.02 (when the word "better" is excluded). Consistent with the conclusion of no evidence of an effect according to the pre-registered significance level α = 0.01, the p-

Table 5. For the positive and negative emotion categories, Linguistic Inquiry and Word Count (LIWC) words most frequently categorized as either positive or negative emotion in the confirmatory dataset.

	Top 20 most frequent words				
Positive emotion	like	good	love	better	pretty
	best	beautiful	important	great	truth
	well	hilarious	kind	amazing	sure
	save	funny	awesome	care	super
Negative emotion	wrong	bad	problem	awful	hate
	war	worst	fight	scary	hell
	worse	sad	lost	afraid	shocking
	weird	seriously	mad	crazy	horrible

values range between p = 3.64x10-2 (corresponding to the variation when the word "pretty" is excluded) and p = 2.95x10-1 (when the word "better" is excluded).

Similarly, regarding negative-emotion words (H2b), in the main confirmatory analyses, we found evidence in favor of the hypothesis that their presence is positively associated with headline success ($\beta = 0.18$, $p = 1.61 \times 10^{-14}$). We found that this conclusion is robust to the exclusion of any specific word from the dictionary, with the effect estimate ranging between $\beta = 0.17$ (corresponding to the variation when the word "wrong" is excluded) and $\beta = 0.20$ (when the word "fight" is excluded). Consistent with the conclusion of a positive effect according to the pre-registered significance level $\alpha = 0.01$, the p-values range between p = 1.60x10-12 (corresponding to the variation when the word "wrong" is excluded) and p = 5.44x10-17 (when the word "fight" is excluded).

Therefore, we conclude that the misclassification of any individual word (as carrying an emotion although it does not) is unlikely to skew the estimated effects and alter the conclusions of our study.

Discussion

Predictability of success

Our analyses are consistent with the hypothesis that the more successful headline in a controlled pair in the Confirmatory Dataset can be predicted at a statistically significant level based on linguistic features alone (H1). Even though the accuracy is significantly higher compared to the random baseline, we note that it is still low, i.e., 54% for all pairs, and less than 60% even for the pairs with the largest difference in the actual click-through rate. When comparing confirmatory to exploratory pairs, even though the number of modeled pairs increased by more than four times, the accuracy of the model fitted on confirmatory pairs did not increase significantly. The fact that predictability did not increase with sample size implies that identifying better headlines based on linguistic features is an inherently hard problem, not merely a sample size issue.

Linguistic hypotheses

Overall, we found that the pilot results generalize to the Confirmatory Dataset. Table 3 summarizes the meta-analysis of the outcomes of the tested linguistic hypotheses on the confirmatory dataset, compared to the exploratory dataset. Out of the eleven tested linguistic hypotheses, we reach the same conclusion with the confirmatory as with the exploratory headline pairs for eight hypotheses. With the p<0.001 significance level, there are no false positives in the pilot analyses, i.e., all significant features on the exploratory dataset are significant on the confirmatory dataset as well. With the p<0.001 significance level, there are three false negatives in the pilot analyses (first-person plural, third-person plural, and indefinite article), indicating that these effects could be detected with confirmatory pairs, likely due to the increased sample size.

We also note that the point estimates from the confirmatory data fall within the 95% confidence intervals of the pilot estimates. This speaks in favor of the robustness of the analyses of the A/B tests since the exploratory estimates generalize to the confirmatory estimates made with unseen data.

Overall, our findings are aligned with recent research [59] using a similar corpus of headline A/B tests run by news publishers. Hagar et al. developed a machine-learned model to predict headline testing outcomes and found that the model exhibits modest performance above baseline, thus concluding that any particular headline writing approach has only a marginal impact. Our findings echo these insights. Similarly, our findings regarding the role of positive

and negative emotion are consistent. While the work of Hagar et al. focuses on developing and evaluating a machine-learned model to establish the predictability of news headline tests and takes the route of predictive modeling, we instead perform explanatory modeling by designing an empirical analysis that relies on the A/B tests.

Thus, our work contributes findings regarding linguistic features that confirm and refine insights derived by Hagar et al. in a different setting (Chartbeat analytics service provider vs. Upworthy publisher) and using a different methodology (predictive modeling vs. explanatory modeling). The discrepancies between certain linguistic cues (e.g., regarding headline length) open the door for future investigations in order to elucidate the precise mechanisms at play.

Understanding the linguistic features that promote information sharing has broad implications, and it can shed light on what compels people to view and share online content. Knowledge of the linguistic features that cause success could be used by benevolent and malevolent actors alike. Benevolent actors might aim to optimize linguistic features in order to maximize engagement in high-stakes contexts such as public health messaging [60,61]. Malevolent actors, on the other hand, might aim to design clickbait [62] and to optimize the linguistic features by tapping into curiosity as the driving mechanisms [63–65].

Of particular relevance for headline design, our confirmatory analyses found evidence supporting the hypothesis that the presence of negative-emotion words is positively associated with headline success. Previous work has found that online systems might promote negativity by design since negative emotions are more important when conveying the meaning of the message with a few words [20]. Our analysis implies that the very format of a short article headline might favour negativity in unintended ways. This finding has implications for the design of online publishing. Future work should aim to understand how the design of our socio-technical systems might promote negativity and what psychological mechanisms might explain why negativity is positively associated with headline success.

Limitations and future work

We believe that our work goes beyond prior research due to the way the A/B test nature of the experiments lets us control content and author confounds, reader confounds, and context confounds. Content and author confounds are controlled since headlines are about the same article with the same content and are written by the same author, reader confounds are controlled since readers are randomly assigned with a headline version without user-specific selection biases, and context confounds are controlled since the image thumbnail and the rest of the website the user landed on were exactly the same. However, when interpreting our results, certain limitations should be taken into account.

First, we acknowledge that we do not remove all potential confounds. The compared headlines are not directly manipulated by the website on a single linguistic variable while keeping the rest of the headline fixed (in Table 6, we provide information about what fraction of comparison pairs differed on each variable). We argue that, even though multiple variables are manipulated simultaneously, we are able to estimate the effects in a regression framework. In order to assess the potential for remaining confounding, we performed correlational analyses (Fig 5) on our design matrix of predictors. Since the correlations are small (ranging between -0.13 and +0.12) and the features are expressed on the same scale, by entering the eleven linguistic variables in a well-specified regression model, we argue that we can interpret effect sizes as causal effects. The advantage of the fact that the headlines are not directly manipulated by the website on a single linguistic variable is that the headlines are naturally written in their entirety as they are. Direct manipulation by insertion or deletion of words would provide a tight control, but might result in unnatural headlines. Future work should determine whether

Table 6. Linguistic variables and the frequency in which they are manipulated among the comparison pairs, in the exploratory dataset, and in the confirmatory dataset.

Linguistic variable	Exploratory dataset: Percentage of comparison pairs differing with respect to this linguistic variable	Confirmatory dataset: Percentage of comparison pairs differing with respect to this linguistic variable
first-person singular	16.3%	16.9%
first-person plural	14.7%	14.3%
second-person	27.8%	27.8%
third-person singular	31.0%	30.4%
third-person plural	17.8%	17.7%
indefinite article	37.4%	37.0%
definite article	37.9%	37.7%
positive emotion	38.4%	37.5%
negative emotion	30.6%	30.9%
number of characters	100%	100%
Flesch reading-ease score	100%	100%

https://doi.org/10.1371/journal.pone.0281682.t006

the results presented here hold when linguistic variables are directly manipulated. Furthermore, future work should examine the ways how linguistic features interact, potentially by analyzing larger numbers of tested headline variants.

Second, we note that Upworthy headlines are written in a specific style. For instance, headlines omit articles by stylistic convention. Therefore, findings from other media might not easily transfer. Future work should determine the extent to which these findings generalize to other domains, including frequently studied literature and social media domains, based on which the hypotheses were developed. Similarly, linguistic features of news articles and their impact on headline success might vary across cultures and languages. Future work should determine the extent to which our findings generalize to other languages beyond English.

Third, the textual cues might not necessarily always match the psychological interpretation. For example, regarding measuring emotions, advanced machine learning approaches to classify the emotion in the headlines exist. Similarly, the combined use of more sophisticated text representation tools, including other dictionary-based approaches and machine learning methods, could be used in future research to increase the predictability of success. However, we opt for using the LIWC dictionary approach due to its simplicity, interpretability, and frequent use in the previous literature. Additionally, the dictionary approach lets us identify the words that carry positive or negative emotion, which is compatible with our dictionary-based approaches of measuring other linguistic features of interest. Similarly, the usage of definite and indefinite articles is an imperfect proxy for generality and specificity of a headline, and headline length may not accurately reflect the quantity of information present in the headline. More detailed analyses are necessary to disentangle headline length, the amount of information that the headline contains, and its conciseness.

Finally, we note that unfortunately no reliable statistics about the website's user demographics are available. We know that in the period when the tests were conducted (between
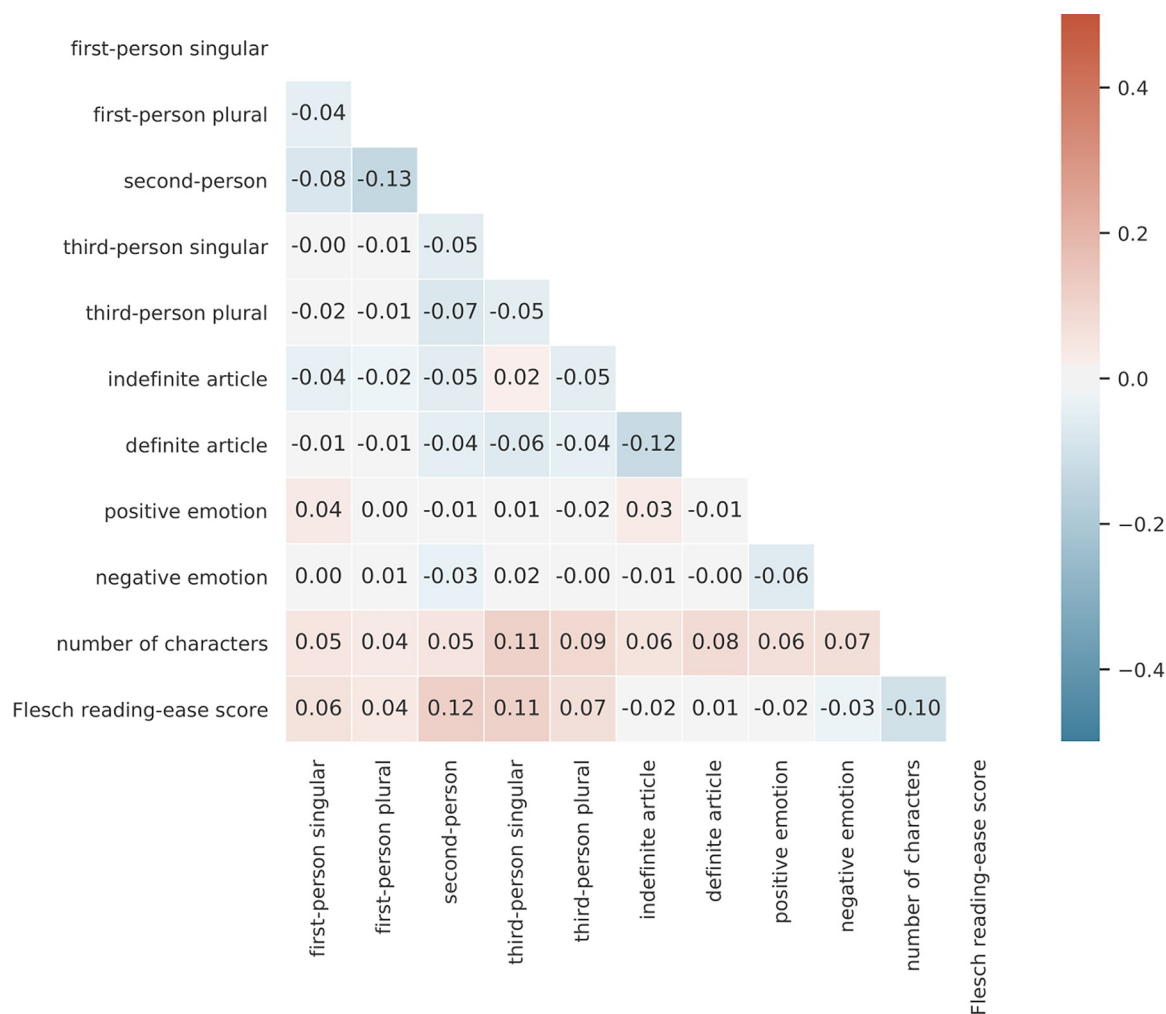
Fig 5. Correlations between linguistic features. Pearson correlation coefficient between values of linguistic features. All correlations are small, ranging between -0.13 and +0.12.

2013 and 2015), Upworthy was a fairly popular website and an important actor in the online ecosystem [66]. Therefore, due to the large user base that it attracted, understanding the features of content that lead to clicks is important, regardless of the demographics. While the results may speak to a biased subset of the overall population, the demographic features of users cannot impact the estimated linguistic effects due to random assignment of headlines. Lastly, during the data-collection period (January 2013—April 2015), socially notable events took place, captured public attention, and were covered in the news. Although we expect the tested linguistic features and their effects on headline success to be universal within the studied context, when interpreting our findings, one should be mindful of the timeframe and social circumstances in which the headlines were tested. Future work should understand the extent to which our findings generalize to other timeframes and other social and historical contexts.

Conclusion

We found that generality (H5a), the use of negative emotion words (H2b), headline length (H3), and the use of first-person singular and third-person pronouns are positively associated

(H6a, H8), while the use of first-person plural pronouns is negatively associated, with success (H6b). Although Upworthy headlines have a specific style, we expect that the psychological processes leading to a click on Upworthy, e.g., referring to general contexts (H5a) or the use of negative emotions words (H2b), can be expected to generalize to different types of content. We believe these insights about linguistic properties that lead to clicks on Upworthy will be of interest to content creators across domains.

Acknowledgments

Author Contributions

Conceptualization: Kristina Gligorić, Robert West, Ashton Anderson.

Formal analysis: Kristina Gligorić, George Lifchits.

Funding acquisition: Robert West, Ashton Anderson.

Investigation: George Lifchits, Ashton Anderson.

Methodology: Kristina Gligorić, George Lifchits, Robert West, Ashton Anderson.

Project administration: Kristina Gligorić, Robert West, Ashton Anderson.

Resources: Robert West, Ashton Anderson.

Software: Kristina Gligorić, George Lifchits.

Supervision: Robert West, Ashton Anderson.

Validation: Kristina Gligorić, George Lifchits.

Visualization: Kristina Gligorić, George Lifchits.

Writing – original draft: Kristina Gligorić, George Lifchits.

Writing – review & editing: Kristina Gligorić, George Lifchits, Robert West, Ashton Anderson.

References

1. Hermida A., Fletcher F., Korell D. & Logan D. Share, Like, Recommend: Decoding the Social Media News Consumer. Journalism Studies 13 (Oct. 1, 2012).

2. Shearer E. & Mitchell A. News Use Across Social Media Platforms in 2020 Pew Research Center's Journalism Project. https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/.

3. Somaiya R. How Facebook Is Changing the Way Its Users Consume Journalism. The New York Times. Business. http://nyti.ms/1yDILEP (Oct. 26, 2014).

4. Carr N. The Great Unbundling: Newspapers & the Net Britannica Blog. http://blogs.britannica.com/2008/04/the-great-unbundling-newspapers-the-net/.

5. Tandoc E. C. Jr. Why Web Analytics Click. Journalism Studies 16 (Nov. 2, 2015).

6. Hagar N. & Diakopoulos N. Optimizing Content with A/B Headline Testing: Changing Newsroom Practices. Media and Communication 7 (1 Feb. 19, 2019).

7. Kohavi R. et al. Online Controlled Experiments at Large Scale. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery, New York, NY, USA, Aug. 11, 2013).

8. Berger J. A. & Milkman K. L. What Makes Online Content Viral? Journal of Marketing Research 49 (2012).

9. Tatar A. et al. Predicting the Popularity of Online Articles Based on User Comments. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics (Association for Computing Machinery, New York, NY, USA, May 25, 2011).

10. Tatar A., Antoniadis P., de Amorim M. D. & Fdida S. From Popularity Prediction to Ranking Online News. Social Network Analysis and Mining 4 (Feb. 12, 2014).

11. Keneshloo, Y., Wang, S., Han, E.-H. & Ramakrishnan, N. in Proceedings of the 2016 SIAM International Conference on Data Mining (SDM) (Society for Industrial and Applied Mathematics, June 30, 2016).

12. Kuiken J., Schuth A., Spitters M. & Marx M. Effective Headlines of Newspaper Articles in a Digital Environment. Digital Journalism 5 (Nov. 26, 2017).

13. Piotrkowicz, A., Dimitrova, V., Otterbacher, J. & Markert, K. Headlines Matter: Using Headlines to Predict the Popularity of News Articles on Twitter and Facebook. In Eleventh International AAAI Conference on Web and Social Media Eleventh International AAAI Conference on Web and Social Media (May 3, 2017). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15657.

14. Hardt, D., Hovy, D. & Lamprinidis, S. 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 EMNLP 2018 (Brussels, Belgium, 2018).

15. Liao Y., Wang S., Han E.-H., Lee J. & Lee D. Characterization and Early Detection of Evergreen News Articles. in Machine Learning and Knowledge Discovery in Databases(eds Brefeld, U.et al.) (Springer International Publishing, Cham, 2020).

16. Tan C., Lee L. & Pang B. The Effect of Wording on Message Propagation: Topic-and Author-Controlled Natural Experiments on Twitter. In Proceedings of ACL (2014).

17. Ferrara E. & Yang Z. Quantifying the Effect of Sentiment on Information Diffusion in Social Media. PeerJ Computer Science 1 (Sept. 30, 2015).

18. Brady W. J., Wills J. A., Jost J. T., Tucker J. A. & Van Bavel J. J. Emotion Shapes the Diffusion of Moralized Content in Social Networks. Proceedings of the National Academy of Sciences 114 (2017). https://doi.org/10.1073/pnas.1618923114 PMID: 28652356

19. Gligoric, K., Anderson, A. & West, R. How Constraints Affect Content: The Case of Twitter's Switch from 140 to 280 Characters. Proceedings of the International AAAI Conference on Web and Social Media 12. https://ojs.aaai.org/index.php/ICWSM/article/view/15079 (1 June 15, 2018).

20. Gligoric K., Anderson A. & West R. Causal Effects of Brevity on Style and Success in Social Media. Proceedings of the ACM on Human-Computer Interaction 3 (CSCW Nov. 7, 2019).

21. Lakkaraju, H., McAuley, J. & Leskovec, J. What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media. In Seventh International AAAI Conference on Weblogs and Social Media (June 28,19 2013). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6085.

22. Guerini, M., Pepe, A. & Lepri, B. Do Linguistic Style and Readability of Scientific Abstracts Affect Their Virality? In Sixth International AAAI Conference on Weblogs and Social Media (May 20, 2012). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4618.

23. Ashok, V. G., Feng, S. & Choi, Y. Success with Style: Using Writing Style to Predict the Success of Novels. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2013).

24. Danescu-Niculescu-Mizil C., Cheng J., Kleinberg J. & Lee L. You Had Me at Hello: How Phrasing Affects Memorability in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers—Volume 1(Association for Computational Linguistics, Jeju Island, Korea, July 8, 2012).

25. Rosen S. The Economics of Superstars. American Economic Review 71 (Dec. 1981).

26. Salganik M. J., Dodds P. S. & Watts D. J. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. Science 311 (Feb. 10, 2006).

27. Martin, T., Hofman, J. M., Sharma, A., Anderson, A. & Watts, D. J. Exploring Limits to Prediction in Complex Social Systems in Proceedings of the 25th International Conference on World Wide Web WWW '16 (ACM Press, New York, New York, USA, 2016).

28. Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report Protocol). Gligorić K, Lifchits G, West R, Anderson A (2021). PLOS ONE 16(9): e0257091.

29. Matias, J. N., Munger, K. & Morris, A. The Upworthy Research Archive: A Time Series of Experiments in U.S. Advocacy Conference on Digital Experimentation MIT. Nov. 2, 2019. https://osf.io/q8g6w/.20

30. Karpf D. Analytic Activism: Digital Listening and the New Political Strategy http://oxford. universitypressscholarship.com/view/10.1093/acprof:oso/9780190266127.001.0001/acprof-9780190266127 (Oxford University Press, 2017).

31. Kamenetz A. How Upworthy Used Emotional Data To Become The Fastest Growing Media Site of All Time Fast Company. https://www.fastcompany.com/3012649/how-upworthy-used-emotional-data-to-become-the-fastest-growing-media-site-of-all-time.

32. Fitts A. S. The King of Content Columbia Journalism Review. https://www.cjr.org/feature/the_king_of_content.php.

33. Matias, J. N. & Munger, K. The Upworthy Research Archive: A Time Series of 32,488 Experiments in U. S. Advocacy in CODE 2019 Conference on Digital Experimentation (MIT, Sept. 9, 2019). https://osf.io/246yq/.

34. Matias J. N. Data in the Upworthy Research Archive The Upworthy Research Archive. https://upworthy.natematias.com/about-the-archive.html.

35. De Vany A. Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry (Routledge, London, 2004).

36. Salganik M. J.et al. Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration. Proceedings of the National Academy of Sciences (Mar. 25, 2020). https://doi.org/10.1073/pnas.1915006117 PMID: 32229555

37. Song C., Qu Z., Blumm N. & Barabasi A.-L. Limits of Predictability in Human Mobility. Science 327 (Feb. 19, 2010). https://doi.org/10.1126/science.1177170 PMID: 20167789

38. Boucher J. & Osgood C. E. The Pollyanna Hypothesis. Journal of Verbal Learning and Verbal Behavior 8 (Feb. 1, 1969). 21

39. Dodds P. S.et al. Human Language Reveals a Universal Positivity Bias. Proceedings of the National Academy of Sciences 112 (Feb. 24, 2015). https://doi.org/10.1073/pnas.1411678112 PMID: 25675475

40. Hu, Y., Talamadupula, K. & Kambhampati, S.Dude, srsly?: The surprisingly formal nature of Twitter's language in Proceedings of the International AAAI Conference on Web and Social Media 7 (2013).

41. Baumeister R. F., Bratslavsky E., Finkenauer C. & Vohs K. D. Bad Is Stronger than Good. Review of General Psychology. https://journals.sagepub.com/doi/10.1037/1089-2680. 5.4.323 (Dec. 1, 2001).

42. Soroka Stuart N. "Good news and bad news: Asymmetric responses to economic information." The journal of Politics 68, no. 2 (2006): 372–385.

43. Trussler Marc, and Soroka Stuart. "Consumer demand for cynical and negative news frames." The International Journal of Press/Politics 19, no. 3 (2014): 360–379.

44. Soroka Stuart, and Stephen McAdams. "News, politics, and negativity." Political communication 32, no. 1 (2015): 1–22.

45. Soroka Stuart, Daku Mark, Dan Hiaeshutter-Rice Lauren Guggenheim, and Pasek Josh. "Negativity and positivity biases in economic news coverage: Traditional versus social media." Communication Research 45, no. 7 (2018): 1078–1098.

46. Grice H. P. Logic and Conversation. Speech Acts (Dec. 12, 1975).

47. Giora R. On the Informativeness Requirement. Journal of Pragmatics 12. https://journals.scholarsportal.info/details/03782166/v12i5-6_s/547_otir.xml (1988).

48. Dor D. On Newspaper Headlines as Relevance Optimizers. Journal of Pragmatics 35 (2003).

49. Simmons, M., Adamic, L. & Adar, E. Memes Online: Extracted, Subtracted, Injected, and Recollected. Proceedings of the International AAAI Conference on Web and Social Media 5. https://ojs.aaai.org/index.php/ICWSM/article/view/14120 (1 July 5, 2011).

50. Eisenstein, J. What to do about bad language on the internet. In Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies (2013).

51. Flesch R. A New Readability Yardstick. Journal of Applied Psychology 32 (1948). https://doi.org/10.1037/h0057532 PMID: 18867058

52. Abbott B. in The Handbook of Pragmatics (John Wiley & Sons, Ltd, 2006). 22

53. Packard G. & Berger J. Thinking of You: How Second-Person Pronouns Shape Cultural Success. Psychological Science 31 (Apr. 1, 2020). https://doi.org/10.1177/0956797620902380 PMID: 32101089

54. Hastie T., Tibshirani R. & Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2nd ed. (Springer, 2008).

55. Wasserman L. All of Statistics: A Concise Course in Statistical Inference (Springer-Verlag, New York,2004).

56. Pennebaker J. W., Booth R. J. & Francis M. E. Linguistic Inquiry and Word Count: LIWC 2007.

57. Aggarwal Chaitanya S. B. Textstat: Calculate Statistical Features from Text version 0.7.0. Nov. 22, 2020. https://github.com/shivam5992/textstat.

58. Fast E, Chen B, Bernstein MS. Empath: Understanding topic signals in large-scale text. In Proceedings of the 2016 CHI conference on human factors in computing systems 2016 May 7 (pp. 4647–4657).

59. Hagar N, Diakopoulos N, DeWilde B. Anticipating Attention: On the Predictability of News Headline Tests. Digital Journalism. (Sep 23 2021).

60. Rimer B. K. & Kreuter M. W. Advancing Tailored Health Communication: A Persuasion and Message Effects Perspective. Journal of Communication 56 (2006).

61. Matz S. C., Kosinski M., Nave G. & Stillwell D. J. Psychological Targeting as an Effective Approach to Digital Mass Persuasion. Proceedings of the National Academy of Sciences 114 (Nov. 28, 2017). https://doi.org/10.1073/pnas.1710966114 PMID: 29133409

62. Dan O., Leshkowitz M. & Hassin R. R. On Clickbaits and Evolution: Curiosity from Urge and Interest. Current Opinion in Behavioral Sciences. Curiosity (Explore vs Exploit) 35 (Oct. 1, 2020).

63. Dubey R. & Griffiths T. L. Reconciling Novelty and Complexity Through a Rational Analysis of Curiosity. Psychological Review 127 (2020). https://doi.org/10.1037/rev0000175 PMID: 31868394

64. Lydon-Staley D. M., Zhou D., Blevins A. S., Zurn P. & Bassett D. S. Hunters, Busybodies and the Knowledge Network Building Associated with Deprivation Curiosity. Nature Human Behaviour (Nov. 30, 2020). https://doi.org/10.1038/s41562-020-00985-7 PMID: 33257879

65. Hidi S. & Renninger K. A. The Four-Phase Model of Interest Development. Educational Psychologist 41 (June 1, 2006).

66. Matias J. Nathan and Munger Kevinand Le Quere Marianne Aubin and Ebersole Charles. "The Upworthy Research Archive, a time series of experiments in U.S. media." Nature: Scientific Data. https://doi.org/10.1038/s41597-021-00934-7 PMID: 34341340