PLOS ONE

REGISTERED REPORT PROTOCOL

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

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

This is a Registered Report and may have an associated publication; please check the article page on the journal site for any related articles.

Citation: Gligorić K, Lifchits G, West R, Anderson A (2021) Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report Protocol). PLoS ONE 16(9): e0257091. https://doi.org/10.1371/journal.pone.0257091

Editor: Shiri Lev-Ari, Royal Holloway University of London, UNITED KINGDOM

Received: February 25, 2021

Accepted: August 23, 2021

Published: September 15, 2021

Data Availability Statement: The data underlying the results presented in the study will soon be available at: Matias, J. Nathan and Munger, Kevin and Le Quere, Marianne Aubin and Ebersole, Charles. "The Upworthy Research Archive, a time series of experiments in U.S. media." Nature: Scientific Datasets. DOI: 10.1038/s41597-021-00934-7 Our analysis code is available on the OSF

Abstract

What makes written text appealing? In this registered report protocol, we propose to 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 comparing multiple headline variants for the same news articles. This unique setup allows us to control for factors that can have crucial confounding effects on headline success. Based on prior literature and a pilot partition of the data, we formulate hypotheses about the linguistic features that are associated with statistically superior headlines. We will test our hypotheses on a much larger partition of the data that will become available after the publication of this registered report protocol. Our results will contribute to resolving competing hypotheses about the linguistic features that affect the success of text and will 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.

Understanding the linguistic features that promote information sharing has broad implications. It can shed light on what compels people to view and share online content. Knowledge

(https://osf.io/j3drb/?view_only=0b50f63b30a24335a04e0f6bed7fd6cc).

Funding: This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Swiss National Science Foundation (SNSF), the Microsoft Swiss Joint Research Center, as well as gifts by Google and Facebook to West's lab. The funders provided support in the form of salaries, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: This does not alter our adherence to PLOS ONE policies on sharing data and materials.

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 [8,9]. Malevolent actors, on the other hand, might aim to design clickbait [10] and to optimize the linguistic features by tapping into curiosity and interests as the driving mechanisms [11–13]. There has been substantial scrutiny of the predictors of success in various domains of text production. On our present focus of news, Berger & Milkman [14] 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 [15–21], 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. [22] discovered several linguistic characteristics of tweets that outperformed closely matched alternatives in an observational study. Other studies on Twitter have investigated how sentiment [23], emotion [24], and length [25,26] affect tweet success. Aside from social media, other studies have used linguistic features to predict success in online communities [27], scientific abstracts [28], literature [29], and quotes [30].

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" [31] 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 [32]. 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 [33]. 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 [22]. 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 report, 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

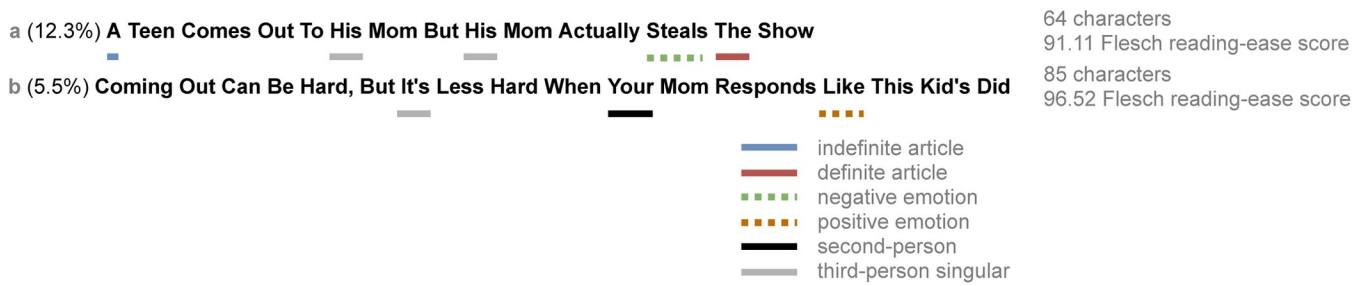

Fig 1. Example of a headline variation experiment pair and derived features. The *a* variant (top) had a higher clickthrough rate (12.3%) than the *b* variant (bottom; 5.5%). The *a* variant contains a definite and an indefinite article, a negative emotion word, and a third-person singular pronoun, whereas 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.0257091.g001

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 [34], 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 [35] and was referred to as "the fastest growing media company in the world" [36,37]. 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 [3,38].

The Upworthy Research Archive contains a total of 32,487 headline variation experiments conducted between January 2013 and April 2015 [39]. 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 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 [38]. 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 is 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. In this registered

report, 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 present registered report protocol. The remainder of the data will be available to researchers whose registered report protocol has been peer-reviewed and accepted. This release process ensures that all proposed hypotheses will be rigorously tested on a 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 analysis.

Within the above-described experimental framework, we will test hypotheses that we developed based on an exploratory analysis of the pilot dataset or that have been proposed as important factors of success by prior literature. Our pre-registered analyses will 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 [40,41]. 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 [32,33,42]. 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 [14] and tweets [22,24]. 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" [43] states that there is a human tendency to use positive language, a result that has been empirically verified across vast and diverse corpora [44]. In the online sphere, positive affect is more prevalent than negative affect on Twitter, indicating that people generally tend to tweet about happy things [45]. 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 formation [46]. For social media, it has been shown that positive content may receive more popularity than negative content [23,27], but negative messages spread faster

Table 1. Design table.

Question	Hypothesis	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 $\alpha = 0.01$.
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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 $\alpha = 0.01$, 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 will be scored on the features, and regression analysis will be performed on all pairs.

https://doi.org/10.1371/journal.pone.0257091.t001

[23], and negative words are more likely to be perceived as relevant to success [26]. Based on these latter results and our findings from pilot data, we form 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 [47]. Meanwhile, the maxim of *relation* may favor brevity, introducing a tension between being informative and being concise [48,49]. 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 [25], and by shortening original tweets to various lengths, Gligorić et al. [26] 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. [50] 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 [22], a hypothesis that is 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 [22]. 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 [51] or the news headlines studied here. In particular, Hu et al. [45] described how stylistic features correlated with readability vary significantly across media. In a study of successful literary works, Ashok et al. [29] 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 [52], 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 [30], whereas the definite article ("the") typically makes reference to something specific and unique [53]. Regarding the appeal of text, Danescu-Niculescu-Mizil et al. [30] found that the usage of general language made movie quotes more likely to be remembered. Similarly, Tan et al. [22] 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 form 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. [29] 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 [28], but were not found to correlate with success in tweets [22]. 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 [20]. 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 [54]. Our pilot analysis yielded a non-significant positive effect of second-person pronouns on headline success. Aligned with previous work, we thus form 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 [22]; they were, however, not found to be associated with popularity in a study of news headlines [20]. 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 hypothesize:

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 is 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 do not propose any exploratory analyses in this 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 [55], 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, it is necessary to 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 [38], 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 [38].

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 consider 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.

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. 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 [56] with a family-wise error rate of $\alpha = 0.05$. Note that in experiments testing more than two comparable headlines, each headline participates in a single

comparison pair. Among such possible matchings of comparable headlines into distinct pairwise comparisons, we select the configuration with the lowest Bonferroni-corrected p-value.

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. To support time-series research, both Confirmatory and Exploratory Datasets are a random sample of A/B tests, stratified by week number. The Confirmatory Dataset is expected to contain roughly four times the number of headline experiments. All hypotheses proposed in this report will be tested on the Confirmatory Dataset using the same Analysis Plan, but with a large sample of unseen data. All data pre-processing steps will be unchanged from what was developed on the Exploratory Dataset.

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) [57] to categorize individual words as possessing either positive or negative emotion. Finally, the *textstat* library for Python [58] is used to compute the Flesch reading-ease score [52] 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 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. In Table 3, we show the fraction of comparison pairs that differ in each of the features.

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.0257091.t002

Table 3. Linguistic variables and the frequency in which they are manipulated among the comparison pairs.

Linguistic variable	Percentage of comparison pairs differing with respect to this linguistic variable
first-person singular	16.3%
first-person plural	14.7%
second-person	27.8%
third-person singular	31.0%
third-person plural	17.8%
indefinite article	37.4%
definite article	37.9%
positive emotion	38.4%
negative emotion	30.6%
number of characters	100%
Flesch reading-ease score	100%

https://doi.org/10.1371/journal.pone.0257091.t003

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. In the registered report to be compiled after analyzing the Confirmatory Dataset, we will say that a hypothesis is supported if its corresponding column in the regression has $p < 0.01$, although for our pilot analyses we considered $p < 0.05$ as preliminary evidence for the hypotheses.

Results on pilot data

The Exploratory Dataset, which is a sample of 4,873 tests stratified by week [38]. As described above, we used the Exploratory Dataset to design research methods and develop the hypotheses presented in this report. A dataset of headline pairs is constructed, features are computed, and a logistic regression model is trained as described in the remainder of this section.

H1: Success prediction. Our first hypothesis posits that our model can predict headline success from linguistic features. Using our regression model, with 0.5 decision threshold on the predicted value, prediction accuracy on the pilot data was 55.23% (*Pearson* $\chi^2(1) = 27.69$, $P<10^{-6}$, *odds ratio* = 1.23, 99% *CI* = [0.534, 0.570]), recall was 55.31%, precision was 55.22%. We conclude with the prediction that our model will identify the better headline from pairs in the Confirmatory Dataset significantly better compared to a random baseline.

H2-H8: Linguistic hypotheses. The remaining hypotheses are assessed from inspecting the fitted regression coefficients (depicted in Fig 2; detailed statistics in Table 4). Our hypotheses (Design Table, Table 1), which posit the presence of an effect and its direction, are based on the Exploratory Dataset results obtained in Fig 2. Overall, we expect these results to generalize to the Confirmatory Dataset.

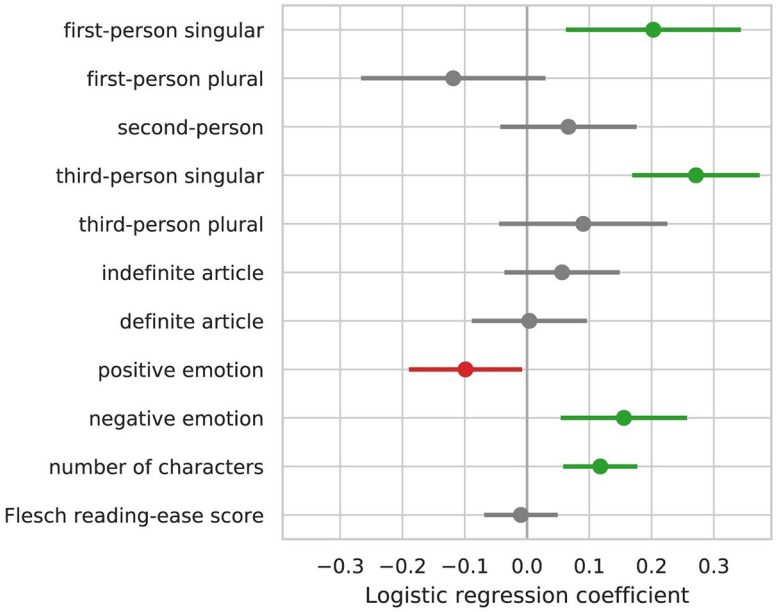

Fig 2. Main regression analysis on the Upworthy Archive Exploratory Dataset. Error bars show 95% confidence intervals on the logistic regression coefficients.

https://doi.org/10.1371/journal.pone.0257091.g002

Table 4. Main analysis regression with Exploratory Dataset. Logit regression analysis for pilot data.

	Dep. Variable:		y	No. Observations:		5048
	Model:		Logit	Df Residuals:		5037
	Method:		MLE	Df Model:		10
	Date:		Sun, 27 Jun 2021	Pseudo R-squ.:		0.01102
	Time:		14:26:22	Log-Likelihood:		-3460.4
	Converged:		True	LL-Null:		-3499.0
	coef	std err	z	P>|z|	[0.025	0.975]
first-person singular	0.1878	0.072	2.613	0.009	0.047	0.329
first-person plural	-0.1228	0.075	-1.628	0.104	-0.271	0.025
second-person	0.0750	0.056	1.348	0.178	-0.034	0.184
third-person singular	0.2611	0.052	5.014	0.000	0.159	0.363
third-person plural	0.1047	0.069	1.525	0.127	-0.030	0.239
indefinite article	0.0486	0.047	1.026	0.305	-0.044	0.141
definite article	0.0047	0.047	0.099	0.921	-0.088	0.097
positive emotion	-0.0910	0.046	-1.969	0.049	-0.182	0.000
negative emotion	0.1540	0.052	2.986	0.003	0.053	0.255
number of characters	0.1209	0.030	3.981	0.000	0.061	0.180
Flesch reading-ease score	-0.0101	0.030	-0.337	0.736	-0.069	0.049

https://doi.org/10.1371/journal.pone.0257091.t004

Acknowledgments

We thank J. Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole for their efforts in creating the Upworthy Research Archive [59]. We are grateful to Upworthy for donating the data for scientific purposes.

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: George Lifchits.

Visualization: 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. Rimer B. K. & Kreuter M. W. Advancing Tailored Health Communication: A Persuasion and Message Effects Perspective. Journal of Communication 56 (2006).
9. 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
10. 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).
11. 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
12. 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
13. Hidi S. & Renninger K. A. The Four-Phase Model of Interest Development. Educational Psychologist 41 (June 1, 2006).
14. Berger J. A. & Milkman K. L. What Makes Online Content Viral? Journal of Marketing Research 49 (2012).

15. 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).

16. 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).

17. 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).

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

19. 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.

20. 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).

21. 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). https://doi.org/10.1007/978-3-030-46150-8_8 PMID: 33103160

22. 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).

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

24. 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

25. 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).

26. 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).

27. 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.

28. 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.

29. 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).

30. 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).

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

32. 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).

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

34. 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.

35. 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).

36. 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.

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

38. 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/.

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

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

41. 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

42. 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

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

44. 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

45. 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).

46. 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).

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

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

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

50. 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).

51. 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).

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

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

54. 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

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

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

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

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

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