Investigating the Relationship Between Keyword Content of Digital Newspaper Headlines and its Effect on Reader Traffic

Annie Huang, Linguistics Department, University of California, Santa Barbara

Abstract. Using data from an online university newspaper, we analyze the amount of information contained in articles that are reflected in their headline and investigate its correlation with the number of page views the articles receive. We utilize computational linguistic methods to pre-process the language data, extract potential keywords and their significances, and compare the amount of significant keywords an article received with its page views. We find that in general, article headlines containing more significant keywords directly related to their articles receive more page views up to a certain length in text. In a digital world with clickbait-titles and misleading headlines, it is important to understand the significance of having article titles that accurately reflect the content that follows in order to maintain trust in readership as well as decrease the spread of misinformation.

1 INTRODUCTION

It has become common for online news sources to utilize ambiguous titles in order to bait the attention of potential readers. These vague titles often don't accurately reflect the actual content of the article, and in many cases are misleading due to the ways they can be interpreted. Similarly, news sources may use words that are relevant to their article content in the headline, but may exaggerate or distort them in a way that is less accurate, but seemingly more sensational. This becomes especially problematic because for most news articles, only the headlines are actually read by readers as they browse through the internet.

In a country where the freedom of the press is legally protected, sources of journalism have an ethical obligation to provide its audience with accurate information. However, it is possible that the use of misleading article headlines may be due to a lack of training in journalists, or the desperate belief that an engaging, albeit inaccurate title that will get the readers' attention is better than a plain but accurate title. In this project, we analyze the shared keywords between the headlines and article content of a student paper at the University of California, Santa Barbara, and explore the relationship between the amount of reflected information and reader engagement with the actual article.

2 BACKGROUND

The Bottom Line is a student-run, student-funded newspaper at the University of California, Santa Barbara. The Bottom Line's readership consists mainly of the undergraduate students of the University of California, Santa Barbara. Because The Bottom Line is funded with student fees, it is expected of the newspaper to deliver accurate information to students in an efficient and honest manner. While the newspaper distributes its physical paper across Isla Vista, California, its digital content can be viewed on its website https://thebottomline.as.ucsb.edu/.

The data of this research project is obtained through The Bottom Line website using WordPress, which the newspaper is hosted on. Each row of the dataset represents one of the 64 articles from the newspaper's 'NEWS' section during the 2020-2021 academic year, with the variables being the article title, written content, and page views.

3 METHODS & RESULTS

Study 1: Headline Length and Page Views.

The first test was to see if there was any relationship between the length of the headline Short news article titles are often vague or too broad, leaving out key information that usually tells more of the specific truths of the article. Therefore, we wanted to see whether the lengths of the titles had any correlation with the page views of the articles.

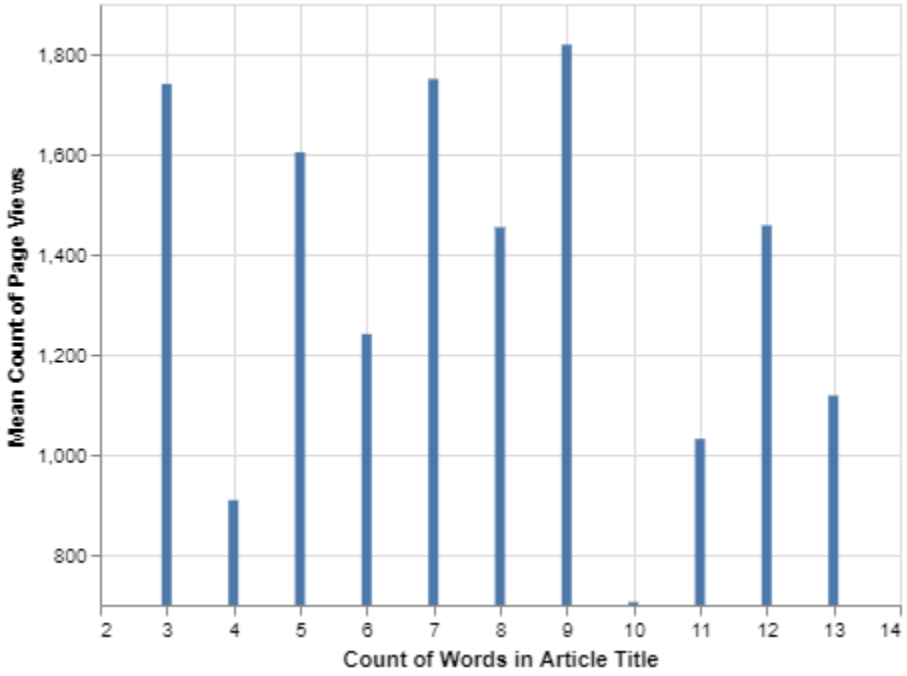

Figure 1: Bar plot representing the number of words in an article title (x-axis) and the average number of page views an article gets (y-axis).

To begin, we measured the word count of the article titles and average count of page views for each word count. From Figure 1, there does seem to be a general increase in page views as the count of words increases, peaking at 9 words and decreasing from there. On average, it seems that article headlines with 3, 5, 7, and 9 words get the highest number of page views.

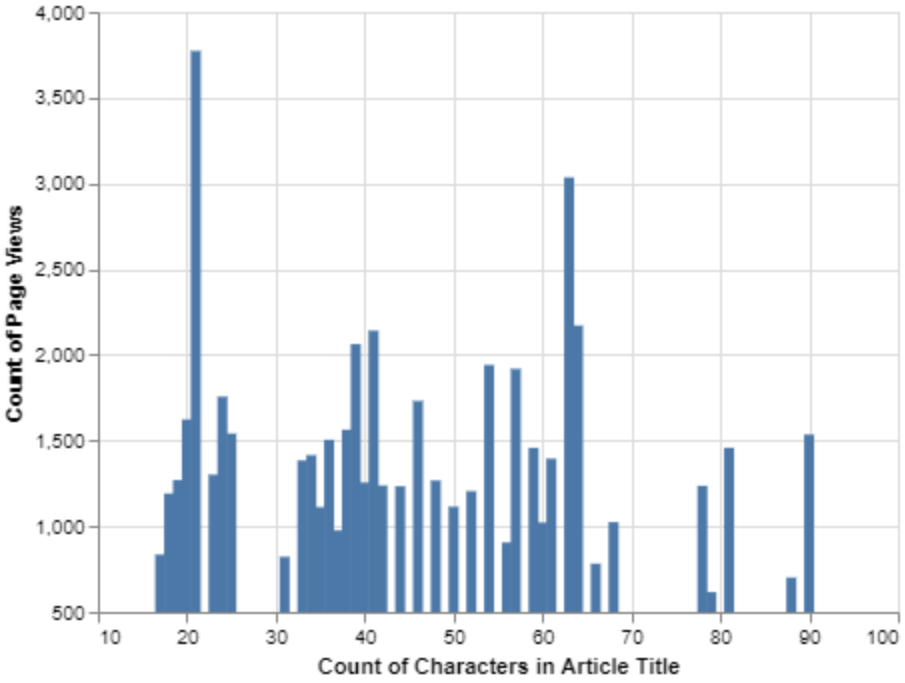

Figure 2: Bar plot representing the number of characters in an article title (x-axis) and the average number of page views an article gets (y-axis).

In addition to the word count in a headline, we also measured the character count in a headline. From Figure 2, we see that the majority of the article title lengths range from 17-70 characters. The highest average number of page views was around 22 characters, with the second highest at around 63. There is no clear trend, but like the plot (Figure 1) with word counts, there is a significant fall in page views as the character counts go over 65.

The data suggests that there may be a positive correlation between the length of an article title and the amount of reader clicks the article receives up to a certain maximum, where the page views begin to decrease with increased headline length instead. It is understandable that journalists must strike a balance between giving enough information to gain the reader's attention, but not make the headline so long that the reader loses focus before even reading the article.

Study 2: Unweighted Shared Headline Keywords and Page Views.

While the methods of study 1 were able to get quick, initial insights on the relationship between the appearance of the amount of information in a headline and the page views, it is unable to recognize the significance of the words or characters. For this reason, an alternate method is necessary in order to factor in the significance of the language used in these headlines and articles. In this study, we do this by taking the keywords of the article content and count how many of these keywords are present in the actual headline. For this situation, the keywords are generated using spaCy's named entity recognition, which is used to gather the entities that are mentioned in the article.

To begin, the article titles and written content are pre-processed using spaCy and are tokenized, lemmatized, and removed of stopwords and punctuation. Then, we use spaCy's named entity recognition

to extract the main entity keywords from each article into a list. Then, we compare the keywords extracted with the keywords in the headline, checking to see how many named entities are named in the actual headline. A keyword in the list of entities is checked only once, meaning that any duplicate occurrences of a keyword in the headline does not factor into the total number of shared entities in a headline. Finally, we use Pearson's Correlation Coefficient to determine whether the number of shared keywords between an article and its headline is positively or negatively correlated with page views.

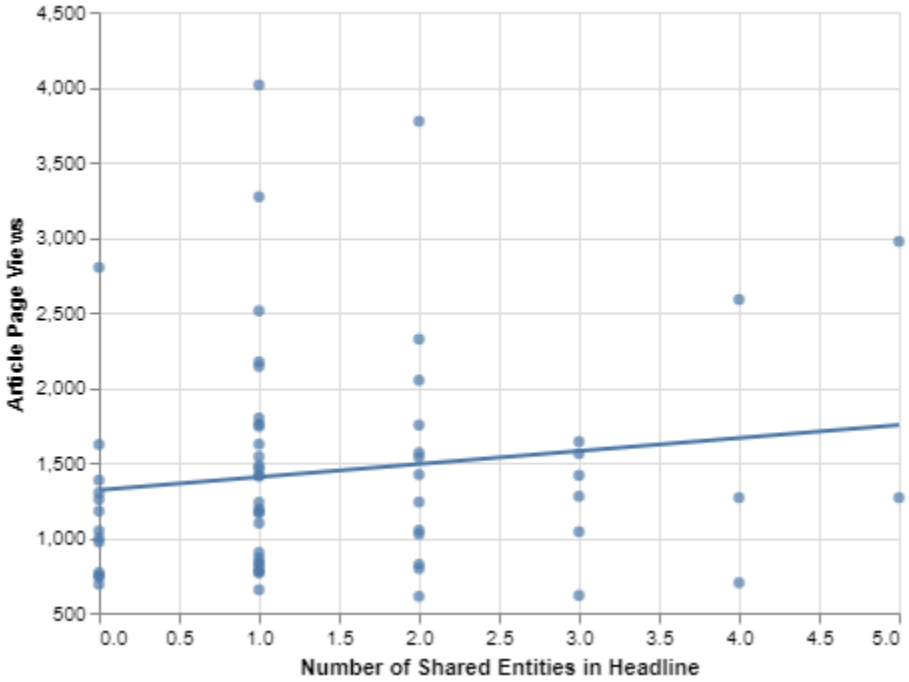

Figure 3: Scatter plot representing the number of keywords (entities) the headline shares with the article content (x-axis) and the number of page views each article gets (y-axis). The blue line represents the line of best fit, or the trend line.

From Figure 3, we see a general positive correlation between the number of shared entities in a headline and article and the number of article page views. We see that the range of page views for headlines with no shared entities is lower than the headlines with 1 or two shared entities. We also see that after 2 shared entities, the number of article page views does not seem to increase nor decrease much, but this may be because of the lack of data regarding headlines with more than 2 shared entities. In calculating the Pearson correlation coefficient, we receive 0.13352267, which suggests a very low positive correlation.

Study 3: Weighted Shared Headline Keywords and Page Views.

The methods of study 2 begin to connect the significant keywords reflected in the content and headlines, but it does not factor in the importance of each keyword. Therefore, in this study we weigh each keyword with its frequency of occurrence in the article and use that frequency in the sum of shared entities in the headline.

Like in the study before, the article titles and written content are pre-processed using spaCy and are tokenized, lemmatized, and removed of stopwords and punctuation. Similarly, we use spaCy's named entity recognition to extract the main entity keywords from each article into a list. Then, we compare the keywords extracted with the keywords in the headline, checking to see how many named entities are named in the actual headline. However in this study, a shared name entity in the headline gets the total number of occurrences in the article added to the total number of shared entities. For example, if an article mentions 'COVID-19' five times and 'COVID-19' is in the article title, then five is added to the number of shared entities in the headline. Finally, we use Pearson's Correlation Coefficient to determine whether the number of shared keywords between an article and its headline is positively or negatively correlated with page views.

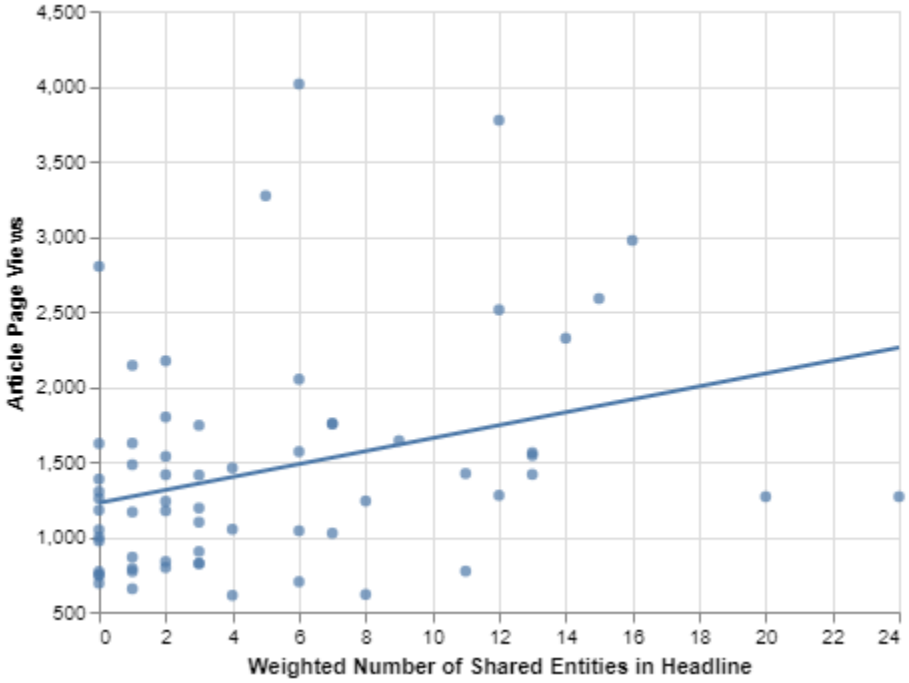

Figure 4: Scatter plot representing the number of keywords (entities) the headline shares with the article content (x-axis) and the number of page views each article gets (y-axis). The blue line represents the line of best fit, or the trend line.

In Figure 4, there is a noticeably steeper positive line of fit between the weighted number of shared entities and the article page views. The majority of this positive correlation occurs within the range of 0-16 weighted shared entities. There are only two data points with a weighted number of shared entities in the headline greater than 16. In calculating the Pearson correlation coefficient, we receive 0.0.32020705, which suggests a low positive correlation.

4 DISCUSSION

Summary of Results

In this paper, we studied the relationship between the number of keywords a newspaper article shares with its headline and the number of page views it received. We saw that there is a slight positive

correlation between the number of words or characters in a headline and the number of page views an article would get. Similarly, we found that article headlines with more keywords relating to the article content also had a positive correlation with the number of page views it would receive, especially if the keywords present in the article headline are words that are repeated and of significance in the article itself.

The findings of this project do suggest that accuracy and relevance of newspaper article headlines to their full article counterparts. It seems that The Bottom Line's readership, namely students of University of California, Santa Barbara, value article headlines that contain a couple significant keywords that are relevant to the article but are less likely to view articles that have no significant keywords or too many. It makes sense that university students would gravitate towards more informative article headlines as they attend a research-based institution, but become disinterested when there is too much overwhelming information.

Limitations

However, the dataset of this study is of a small sample size. The Bottom Line has over 1,700 published articles in the 'NEWS' section, but this dataset only studies the 64 of them that were written and published in the last 9 months. This may be good as it allows the nuances of the articles in the last academic year to shine through. This is because older articles do get more views over time, which could lead to potential biases or bring up inconsistencies. Even so, the dataset is small and may not provide enough information for this study to come up with a conclusive result.

Furthermore, it is difficult to measure whether a headline is misleading or not when it concerns headlines that do use the significant keywords of an article but add other words or formatting in the headline that makes it inaccurate. It is possible that a list of words or a model would have to be created in order to catch any common occurrences of headline exaggerations or sensationalization.

Lastly, the use of the named entity recognition is not enough to capture all the keywords of an article as there are often significant non-named entities that play a large role in the newspaper articles. In addition, using only named entity recognition captures only the entity, not the actions, relationships, or effects of the entity. With only named entity recognition, the full main idea of an article can not be captured or compared to the headline.

Future Research

Despite the shortcomings of this study, the topic of newspaper headline accuracy and its relationship to reader response and viewership should be further investigated for not only marketing in journalism, but for the ethical spread of honest, accurate information. As stated before, this study would benefit largely from a larger sample of articles. Potentially, articles from sections outside of 'NEWS' can be used as well, as there is much more variety and creativity -- as well as sensationalized inaccuracy -- in other newspaper sections, such as opinions. In addition, instead of relying on named entity recognition, better keywords may be extracted through simple frequency analysis across all non-stopwords of an article, or a combination of multiple methods in order to get more words of significance. In fact, even n-grams can be used instead of words. Either way, the results of this study can only be improved from here.