Breaking News

acharle5, fvidhani, msulima2, ipinedad

Hypothesis

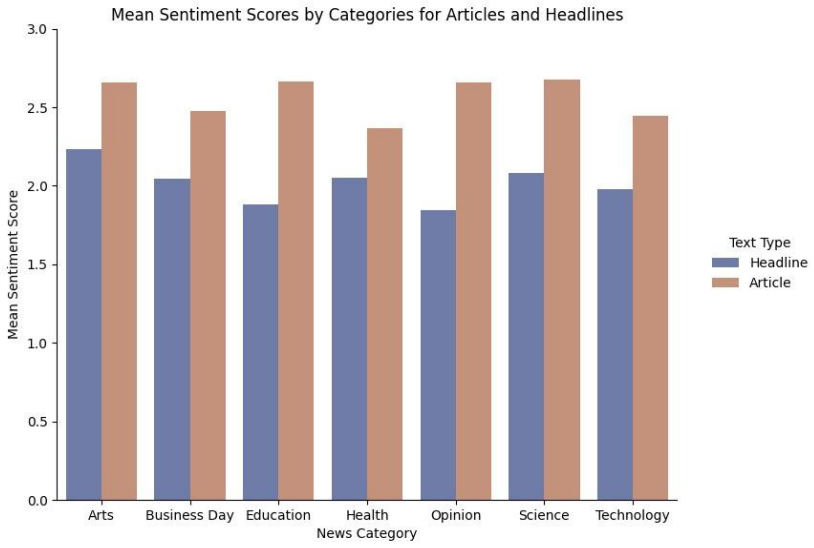
Sensationalism. Spin. Clickbait headlines. Time and time again, it seems that most of the news headlines we read today are imbued with negativity. But are these negative headlines truly reflective of their respective article contents? Or is the media leveraging negativity bias to capture our attention? To draw in more readers? To maximize their profits? By performing sentiment analysis on headlines and their corresponding article content, we investigated our primary hypothesis: article headlines are more negative than their respective article bodies. We also tested our other hypotheses using LSTM-model generated sentiment scores for articles and their headlines.

Data

We scraped the data from the NYT articles database using the NYT Article Search API https://developer.nytimes.com/docs/articlesearch-product/1/overview. The resulting dataset is a collection of articles published between 2018 and 2022. Each article is classified as belonging to one of the following categories: business, technology, science, education, arts, health, and opion. The columns of interest in our dataset were category, headline, and article_text. To clean the data, we converted the values in our columns of interest to solely consist of lowercase letters, stripped the columns of punctuation, removed stop words (commonly used words like “the,” “and”, or “an” that do not add meaning), and performed lemmatization to remove grammar tense. For hypothesis tests that required author names, we removed articles where no author name was provided.

Findings

Claim #1: The mean sentiment scores of headlines is significantly lower than the mean sentiment scores of their corresponding articles.

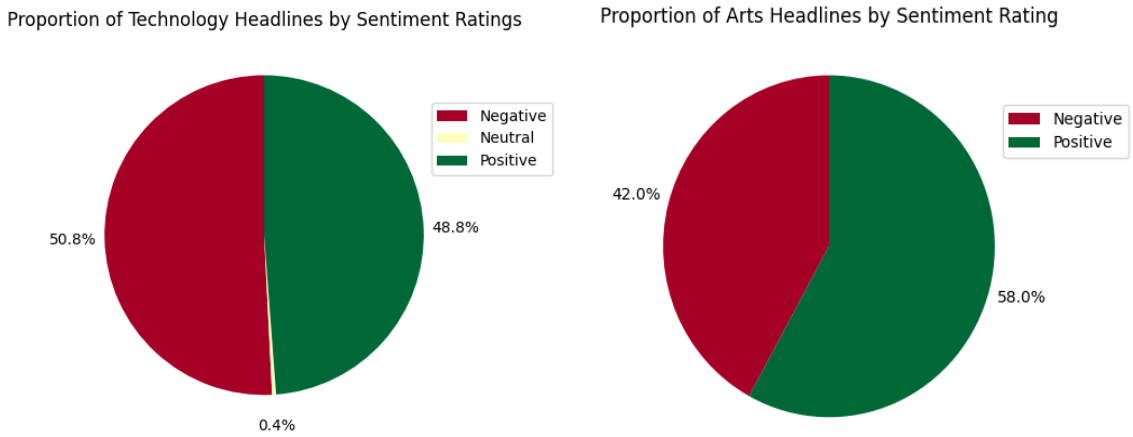


Since each article has two paired measurements (headline rating and article rating), we ran a paired t-test to assess whether headline sentiment ratings were significantly different from article sentiment ratings. Based on the results of our paired t-test, we see that headlines seem to be more negative than their corresponding articles ($t = -15.84$, $p = 1.52e\text{-}52$). Figure 1 illustrates how the mean sentiment scores for articles and headlines differ across the different categories.

Claim #2: The difference in the sentiment scores between headlines and articles is similar across all authors.

We ran a one-way ANOVA test on the computed mean differences between headlines and articles for each author to assess whether certain authors were more likely to write headlines with sentiment ratings that diverge from the sentiment ratings of their corresponding articles. We found that there was no statistically significant difference in the mean differences in sentiment ratings between headlines and articles for the authors in the dataset ($F = 1.04$, $p = 0.30$). While we predicted certain authors to have more divergent sentiment ratings between headlines and articles than others, the results of our one-way ANOVA test suggest otherwise.
….

Claim #3: Arts headlines tend to be more positive than technology headlines.



With the rise of technology, we have been seeing more negative article headlines conveying the dangers of technology. On the contrary, articles about the arts tend to be less negative. We wanted to investigate whether our experiences with arts and technology headlines hold up statistically. The results of our two-sample t-test showed that the mean sentiment score of arts headlines is significantly greater than the mean sentiment score of technology headlines ($t = 2.68$, $p = 0.0038$). Figure 3 illustrates the distribution of sentiment scores for arts and technology headlines.

Breaking News ML Model

Goal

Build and train an LSTM model for sentiment classification of short and long formed text.

Data

Our LSTM model was trained on the SST-5-FINE GRAINED dataset, which contains movie reviews and their corresponding numerical sentiment labels ranging from 0 - 4 (inclusive), corresponding to Strongly Negative, Weakly Negative, Neutral, Weakly Positive, and Strongly Positive sentiments, respectively

Model+Evaluation Setup

Our machine learning model task was to create a sentiment analyzer that outputs a sentiment rating given a body of text. The outputs were integers ranging from [0-4] that represent the sentiments strongly negative, negative, neutral, positive, and strongly positive respectively. With regards to our stated goal, we then used these sentiments to classify news article headlines and text bodies in order to test our hypotheses. Our model uses an LSTM architecture to perform its analysis. The goal of LSTM models in general is to allow the model to remember the important information gleaned from an earlier part of an input whilst considering the latter parts. In the context of our purposes, text classification, this is hugely important because it means that contextual significance of words processed early on within the text body are not forgotten as the model proceeds. This is especially important when considering that article bodies can be extremely lengthy.

Results and Analysis

In the end our model was able to achieve about 39% accuracy. We tried various performance improvements such as dataset stratification, dropout regularization, different hyperparameter configurations, and different model architectures (such as varying the number of LSTM or dense layers). However each optimization had negligible if any impact on the model's performance. As such we decided to remove these optimizations and stick to the simplest model. Although our model accuracy seems low, within the context of the classification task we believe that the results are satisfactory.

Of 26 papers released that have models trained and tested on the SST-5 Fine grained dataset, the worst had an accuracy rate of 41.6% and the best had an accuracy rate of 59.1%. Additionally the aforementioned papers use methods far outside the scope of this class.

As to why even the best models have not been able to achieve better results on this dataset we have a few speculations. One is that there simply isn't enough information to be gleaned from the article text alone. It is possible that important contextual information that we take for granted, such as the world's current political affairs or linguistic idiosyncrasies that ultimately stymy the models' performance. Without access to this information it may even be impossible to achieve a significantly higher accuracy. Another possibility is that the difference between sentiment that are close on the scale (e.g. strongly positive and positive) are to a certain degree arbitrary or person dependent. If the criteria used to label the dataset in the first place is not consistent then the underlying pattern that the model is trying to learn may not even be realizable in the first place. When we changed our criteria of correctness to be a prediction within 1 unit of the true classification our accuracy shot up to 77%. This leads us to believe that our model is correctly recognizing the sentiments of these texts but may not be able to correctly identify the nuances that differentiate a strong sentiment from a weak/normal one.