Analyzing the Sentiment of The New York Times

University of South Florida St. Petersburg
COLLEGE OF ARTS AND SCIENCES

by Luke Daniel Cross

Findings

When headlines lacking polarized words are accounted for, we see a trend line in between comments and headlines with an **R^2 ranging from .02 to .11** and a **consistently significant P-value** for nearly every news section.

Why Sentiment Matters in Journalism

Our era of click bait, fake news and politicized media has cast a critical spotlight onto deceptively structured headlines and how they influence readership. Much of the existing research delves into the implications of phrasing on article headlines -- it is common to find that even minor shifts in the focus of a headline alters what aspects of the article are remembered.

While this is a prime example of the need for conscientious, accurate reporting, it offers little for journalists already taking steps to ensure precise diction and clarity.

Instead of analyzing how a misleading headline *explicitly* alters the way a reader views content, this research attempts to explain how the sentiment inherent to an otherwise fair headline *implicitly* alters a readers reaction to the article.

Even the most objective headline likely uses language rife with implicit sentiment, often out of necessity. For example,

"Senator cuts budget"

is meaningless without context, but negative connotations associated with the word *cut* psychologically prime the reader to assume negative sentiment going into the article.

Using the *New York Times* archives of online articles and their corresponding comments, this theory is tested by passing said comments and headlines through a Naïve Bayes classifier trained on film reviews to quantify sentiment, providing a means of comparing headline sentiment polarity to corresponding comment polarity.

Understanding Sentiment Analysis

Sentiment analysis uses natural language processing to computationally extract and quantify subjective information from unstructured data. The most common application is in marketing to understand customer feedback, but the process can be applied to any natural language data.

The software used for this analysis comes from *TextBlob*, a library of text processing tools built in the programming language Python. We will be using *TextBlob*'s classification and sentiment analysis software, the latter of which utilizes a Naïve Bayes classifier.

To understand Naïve Bayes, one must be familiar with *Bayes' Theorem*,

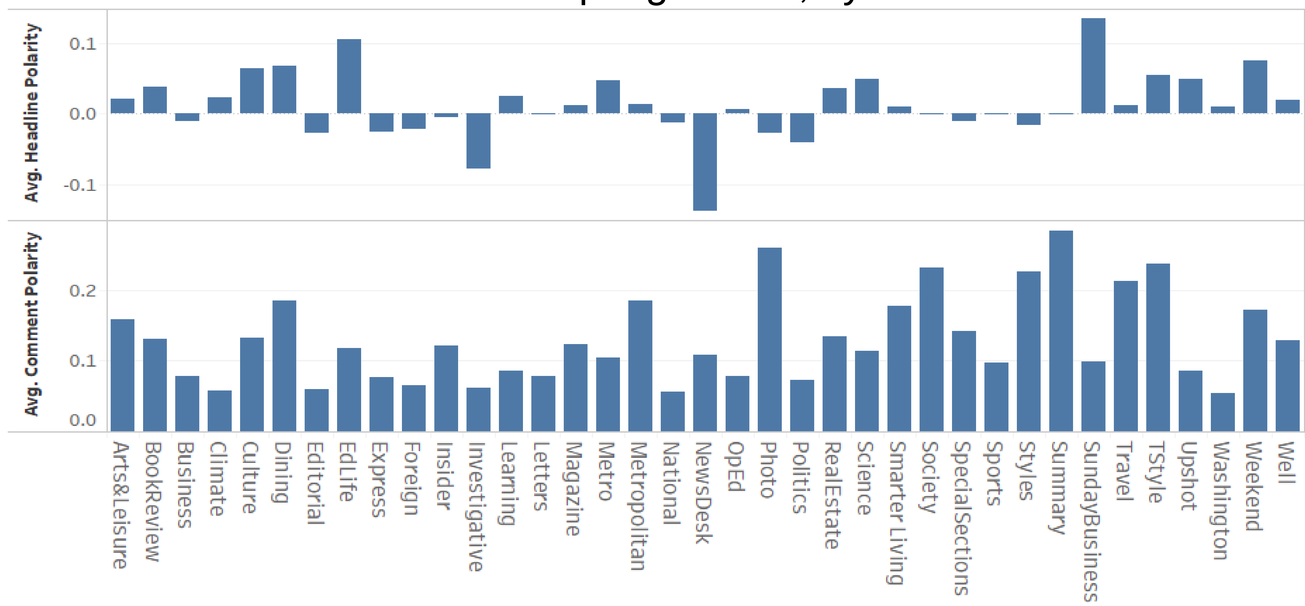
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

which relies on *conditional probability*, the probability of both events A and B occurring is equal to the probability of event A multiplied by the probability of event B happening after event A, all divided by the probability of B.

This classifier, when provided with a dictionary of polarized words and their associated sentiment value from -1 to 1, is applied to two sets of film reviews to "train" its ability to predict[1]. The training set is labeled as positive or negative while the testing set remains unlabeled so that the classifier's conditional probability predictions can be tested and refined.

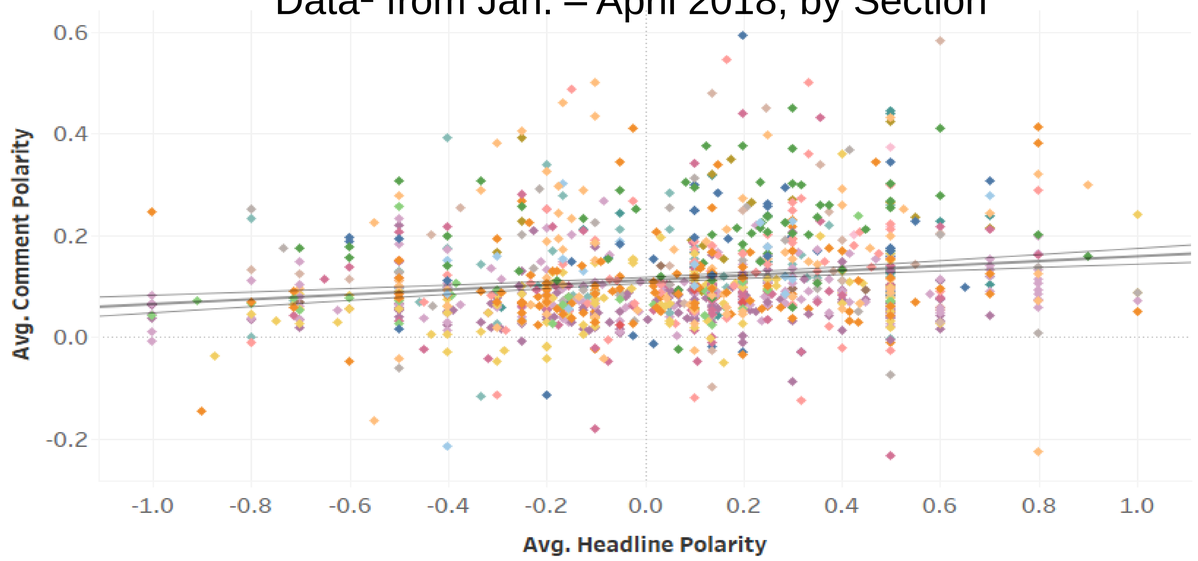
Total Average Headline & Comment Polarity

Data[2] from Spring '17/'18, by Section



Comment and Headline Polarity Correlation

Data[2] from Jan. – April 2018, by Section



Special Thanks to:
[1]Xavier Falco, Rafi Witten & Robin Zhou for training the Stanford NLTK
[2]Aashita Kesarwani for the tools to extract NYT data

For More Info:
lukecross@mail.usf.edu