Clickbait Headline Predictor and Analysis
Capstone 3
Amanda White

Business Problem:

A news site wants to investigate if their headlines have evolved to compete with other new sites who are largely identified as “clickbait”. The news site wants to ensure that they are not contributing to the messy world of misinformation which is often accompanied by clickbait articles that are published to get views, not to keep people informed. The news site wants to ensure they are maintaining their standing as a credible news site.

Data

For this project, I used two headline datasets with published dates. The first data set was headline data from a reputable news site in Australia which represents headlines that are largely not clickbait, the second data set made up the clickbait headlines and was from The Examiner, a news site that is no longer in business and is considered to be a content farm and not a reputable news source.

To begin the project, I merged the two datasets and only kept headlines from 2010-2015 because those were the years both datasets had in common. The data required very little cleaning because there were only two columns: headline text and publish date. I did remove any headlines with less than ten characters because I found some headlines were just one word, or random letters and did not make sense on their own. There was an unbalanced ratio of headlines, there were roughly 3 million clickbait headlines and 400,000 non clickbait headlines. I chose to randomly drop clickbait headlines to balance the data and avoid bias.

Exploratory Data Analysis

I began EDA by comparing the differences in headline length between clickbait and non-clickbait headlines. A histogram of the headline length showed that length of clickbait headlines have a slightly wider spread than non clickbait headlines (figure1).

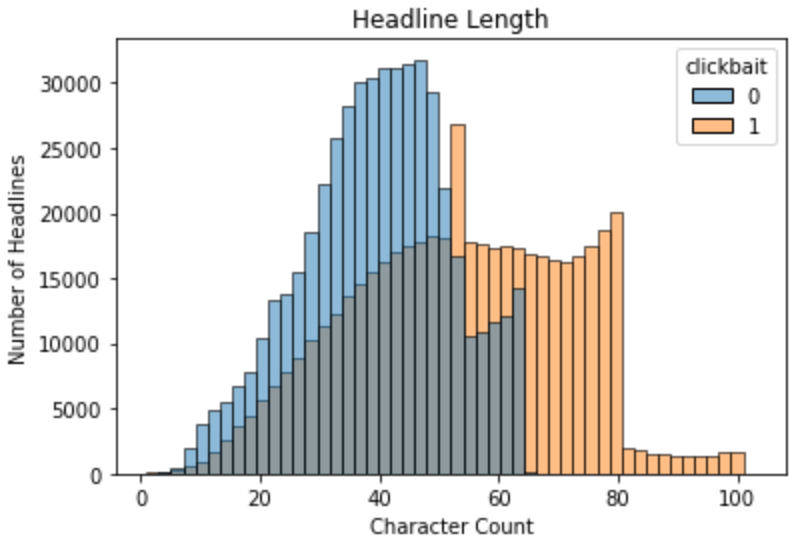

Figure 1. Histogram of headline length. Clickbait headlines are shown in orange and non clickbait headlines are in blue.

The next step was to count how many question marks and exclamation points each type of headline used. None of the non clickbait headlines had either a question mark or exclamation point while 8% of the clickbait headlines had a question mark or an exclamation point. Next, I moved on to counting and removing stop words (it, a, the, and, etc.). There wasn't much difference in the amount of stop words used (figure2).

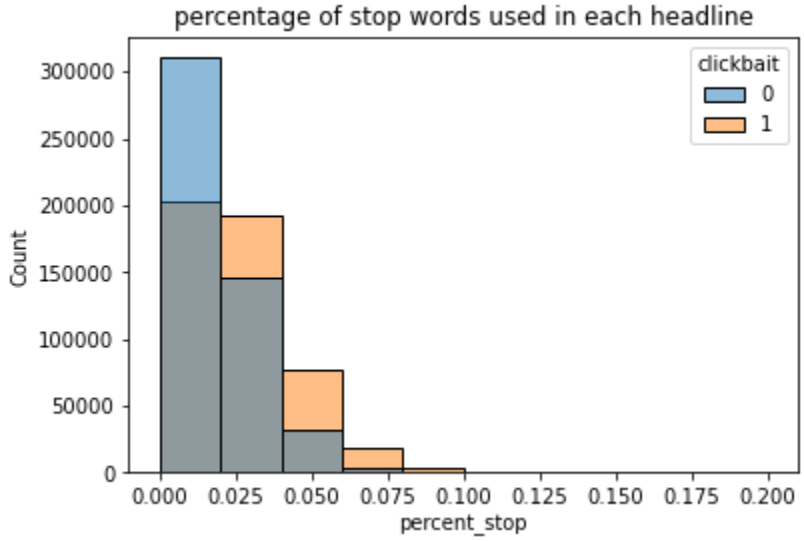

Figure 2. Percentage of stop words used in each headline. Clickbait headlines are in orange and non clickbait headlines are in blue.

I next looked at word frequency. The bar graphs below show the top thirty most commonly used words used in the headlines (figure3). The most notable difference

between word frequency is the presence of numbers and years in the clickbait headlines.

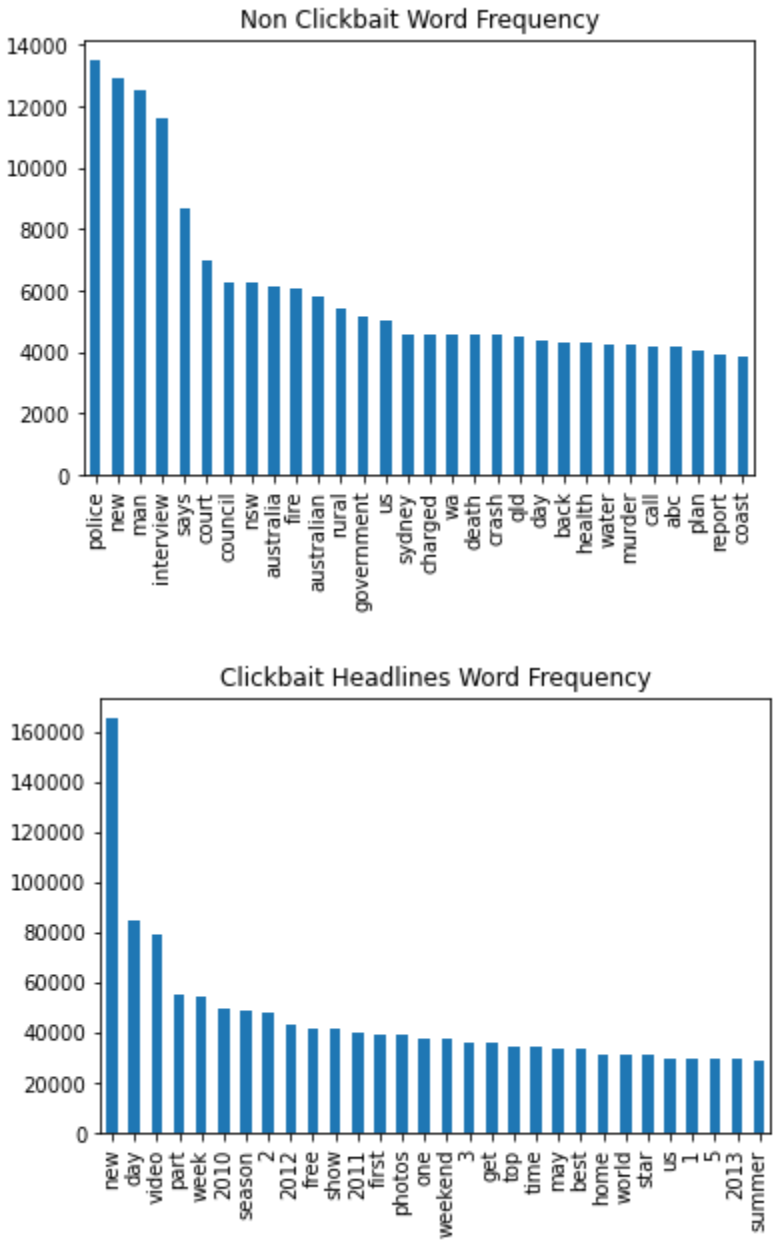

Figure 3. The top graph shows the top thirty most commonly used words in non clickbait headlines and their frequency. The bottom graph shows the same for clickbait headlines.

Preprocessing and Modeling

To prepare the data for modeling, I removed all stop words, punctuation, whitespaces, and numbers. This step ensures that only the key words are left and a model does not make predictions on irrelevant punctuation and stop words that add little to a headline's meaning.

I then split the data into a 75/25 train/test split and tokenized the individual words and converted them to sequences. The next step was to apply an embedding layer which essentially informs the model how words are related to each other. For this project I used GloVe embedding.

For modeling, I used a RNN and ended up with a predictive accuracy of 91%. To use the model, I created a function which can be used with individual headlines. The function preprocesses the data and then runs it through the RNN I created with the training data. If the model scores the headline less than 0.50 it is not clickbait and anything above 0.50 is clickbait. While this model gives a clear yes or no answer, it is to be used with the judgment of the article writer. This should be used as a tool to ensure headlines are not being sensationalized and thought through. If the model determines a headline to be clickbait it is up to the writer to decide if the headline is representative of the article or not.

Future Research

To dive a little deeper into examining if ABC news has been adding to clickbait and sensationalized news, I would like to conduct a cosine similarity analysis. Ideally, I would separate the headlines by the publish year, and compare each year of the non clickbait headlines to clickbait headlines. This would indicate if ABC news has become more like clickbait over time.

To improve model performance, the best way would be to retrain the model using data made up of confirmed clickbait and non clickbait headlines. The model could also be improved to also take the article text into account and so the headline would be determined clickbait or not based on the contents of the associated article.