Sentiment Analysis on Finance Headlines from the New York Times

Abstract

Machine Learning strategies can be applied to learning more about the economy and identify trends in the past that can forecast what is to come in the future. This project seeks to explore how financial news headlines potentially corroborate with economic activity. The project explores how different machine learning algorithms can perform sentiment analysis and then apply that to the New York Times financial news headlines. The analysis then looks at how negative, positive, or neutral headlines are reflective of the current economic swings in 2020 due to the COVID-19 pandemic.

Intro/Background of the Problem

Understanding media and society sentiment about financial health is an important aspect of understanding the potential of economic patterns and forecasts. One of the measures of the public sentiment is in the media headlines. If the positive, negative, or neutral attitudes of headlines regarding the financial world can be measured, there can be a better understanding of how an economy will act. If media sentiment correctly reflects changes or stability for different economic metrics, this could be an important forecasting tool in understanding where the economy is heading. The model could then be used to analyze daily headlines from various newspaper sources to analyze societal sentiment about the economy.

Data Source

To start the project, the primary source of the data will be a dataset of Financial News Headlines and the label of positive, negative, or neutral. The dataset contains 4,846 entries. The dataset will be used to train a machine learning model to predict the sentiment of the headlines. The dataset is available at the link below.

https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news#all-data.csv

For further analysis, the New York Times API will be used to pull headlines and see what sentiment the model applies to them. The documentation for the New York Times API is available at the link below:

https://developer.nytimes.com/

Data Preparation

The dataset being used for testing and training the machine learning models have two columns: sentiment and headlines. The sentiment is the target variable and the headlines are the features that will be used to predict the sentiment.

The sentiment variable has three values: neutral, positive, and negative. While some machine learning algorithms can work with categorical variables, it is important to convert these to numeric so that they can work with any possible machine learning algorithm. Negative is assigned the value of 0, Neutral is assigned the value of 1, and Positive is assigned the value of 2.

The more complex part of the data preparation was getting the headlines ready for text processing. The first step was to remove all of the punctuation from the headlines. The second was to use the NLTK library to tokenize the sentences. The third step was to use the NLTK library of stopwords to remove common words that potentially have little impact on the final sentiment of the sentence. At one point in the preparation, a porter stemmer was applied to convert the words to their root. However, after model evaluation, this appeared to have little impact on the final result.

After this preparation, a headline such as “According to Gran, the company has no plans to move all production to Russia although that is where the company is growing” becomes “‘According', 'Gran', 'company', 'plans', 'move', 'production', 'Russia', 'although', 'company', 'growing'”.

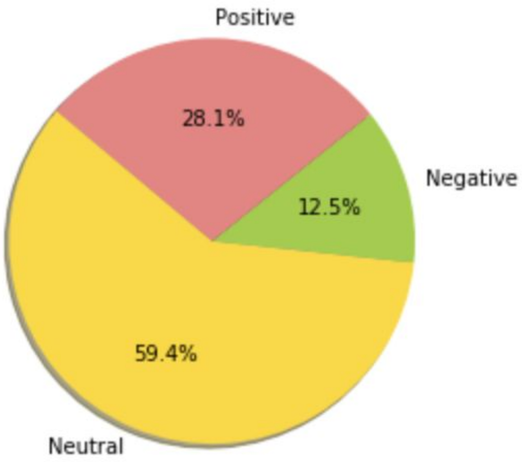
At this point of the data preparation, a Vectorizer was applied to the headline data. For model building and evaluation, both the TF-IDF vectorizer and the CountVectorizer were used to see which one would generate the best results. The differences between the two vectorizers were insignificant and the TF-IDF vectorizer was used to prepare the data for the final model.

There were four parameters added to the vectorizer. The first was an Ngram range of (1,2). This required the vectorizer to consider not only a single word but the word next to that word. For example, without this Ngram range, “good” would be considered a positive word even if it was in the phrase “not good” which would be considered negative. The Ngram range allows for the possibility that a neighboring word may change the sentiment of the word. The second and third parameters were a max frequency of 0.9 and a minimum frequency of 0. The reason for this is that there are some words that have high frequency that have little to no impact on the final sentiment decision (this will be explored more in the Data Exploration section). These parameters focus on words that are frequent enough to be noticed but not so frequent that their relationship to sentiment is weighted heavier than in reality. Finally, the fourth parameter set the maximum features to 4,000. This parameter is in line with the reasons for setting minimum and maximum frequencies.

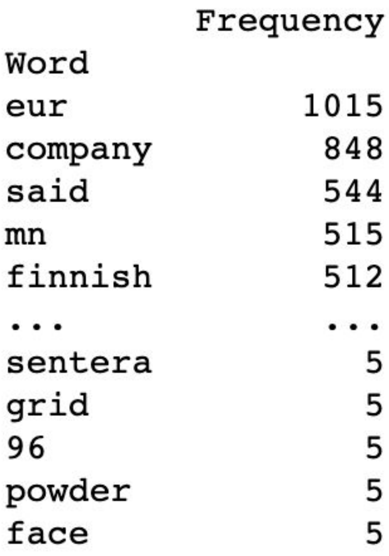
The second part of the overall data preparation was pulling data from the New York Times API. The API request included date parameters so that time could be a part of the final analysis along with the keyword of finance to get the correct category of headlines. The New York Times allows for 10 API requests a minute and 4,000 requests a day. The collected data was then imported into a Pandas dataframe and exported to CSV. For the model, the data went through the same text processing steps as the training dataset.

Data Exploration

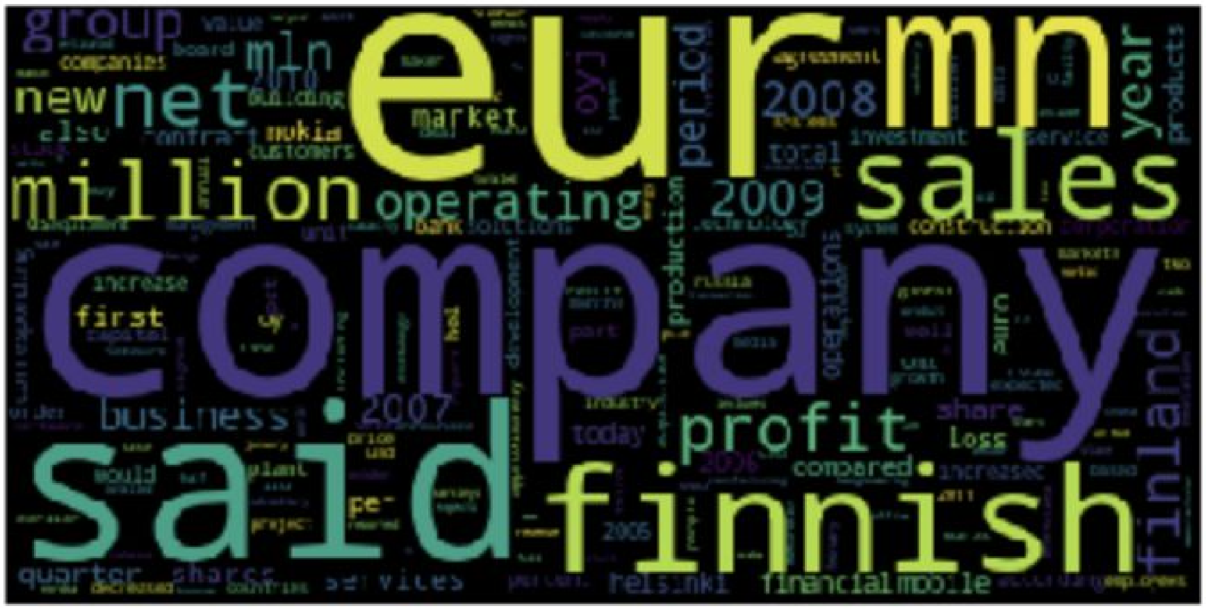
It was important to understand the ratio of the target variable (sentiment) in the data that would be used to build the model. As seen in the pie chart below, the majority of headlines were assigned a neutral sentiment. The headlines with a positive sentiment are almost double in volume to those that are negative. The concern here is if these ratios will impact the model's ability to distinguish between negative and neutral since the ratios for training the model are so different.



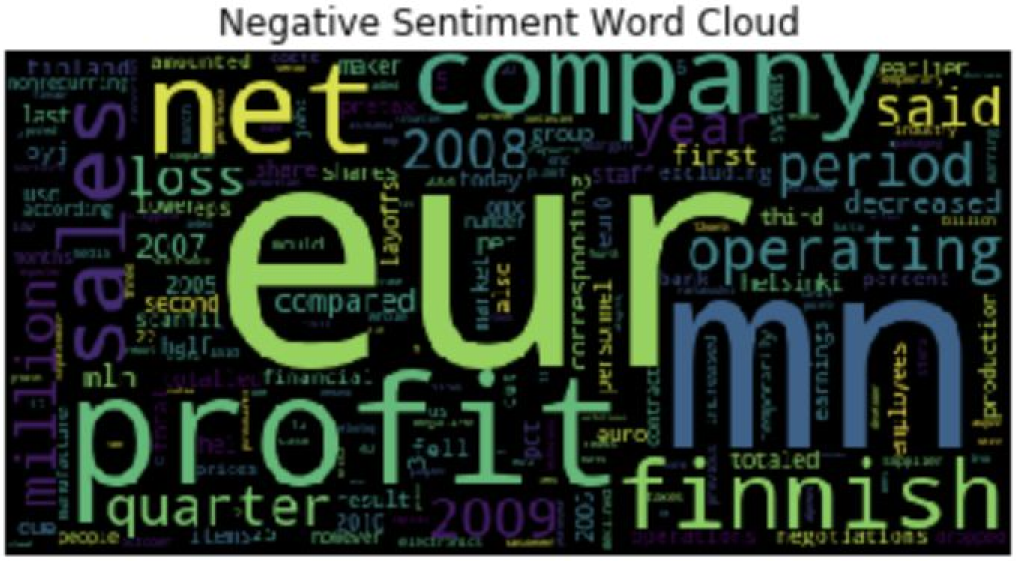
Another important part of understanding the data is understanding the words that are being used for the sentiment analysis and their frequency. Understanding these aspects will inform how the model should be customized for the best results. Two tools were used to help understand these aspects of the data. First, a dataset of the words and their frequency was created (image displayed below). Second, a word cloud was created to visualize the frequency of the words (image displayed below).

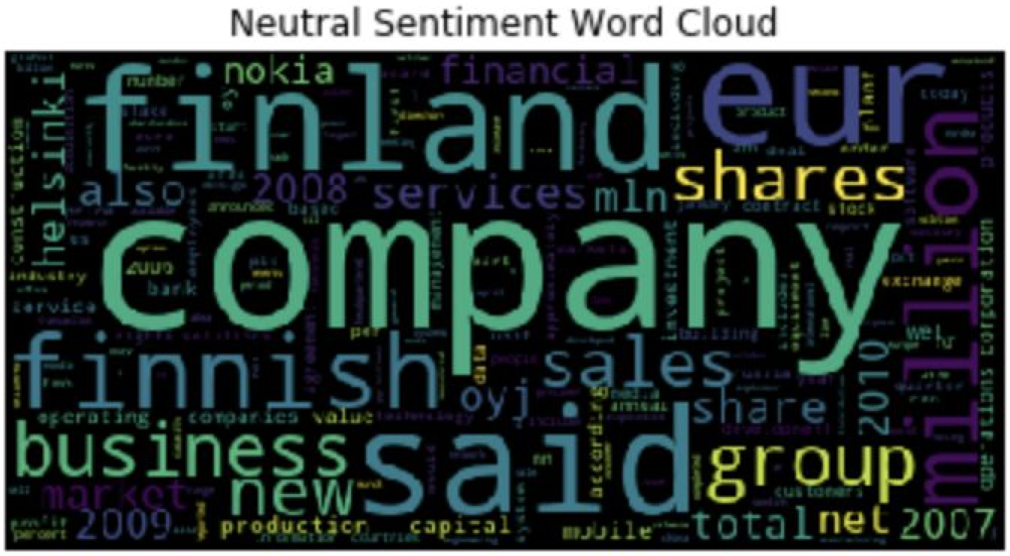


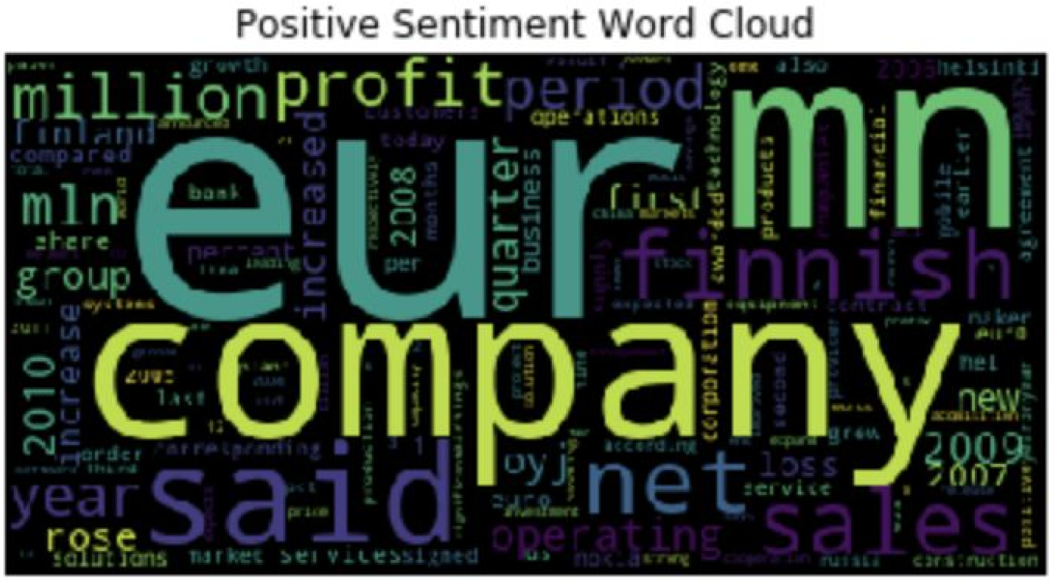
```
         Frequency
Word
eur           1015
company        848
said           544
mn             515
finnish        512
...            ...
sentera          5
grid             5
96               5
powder           5
face             5
```



As noted in the Data Preparation section, a parameter was set so that some of the most frequently used words were not included in the training and test dataset for building the model. Some of the reasons why can be seen in the previous word cloud (eur is the most common word and will likely have little role to play in determining the sentiment of a news headline). For further verification, a word cloud was created for the most common words for headlines labeled positive, negative, and neutral and they are displayed in the images below.







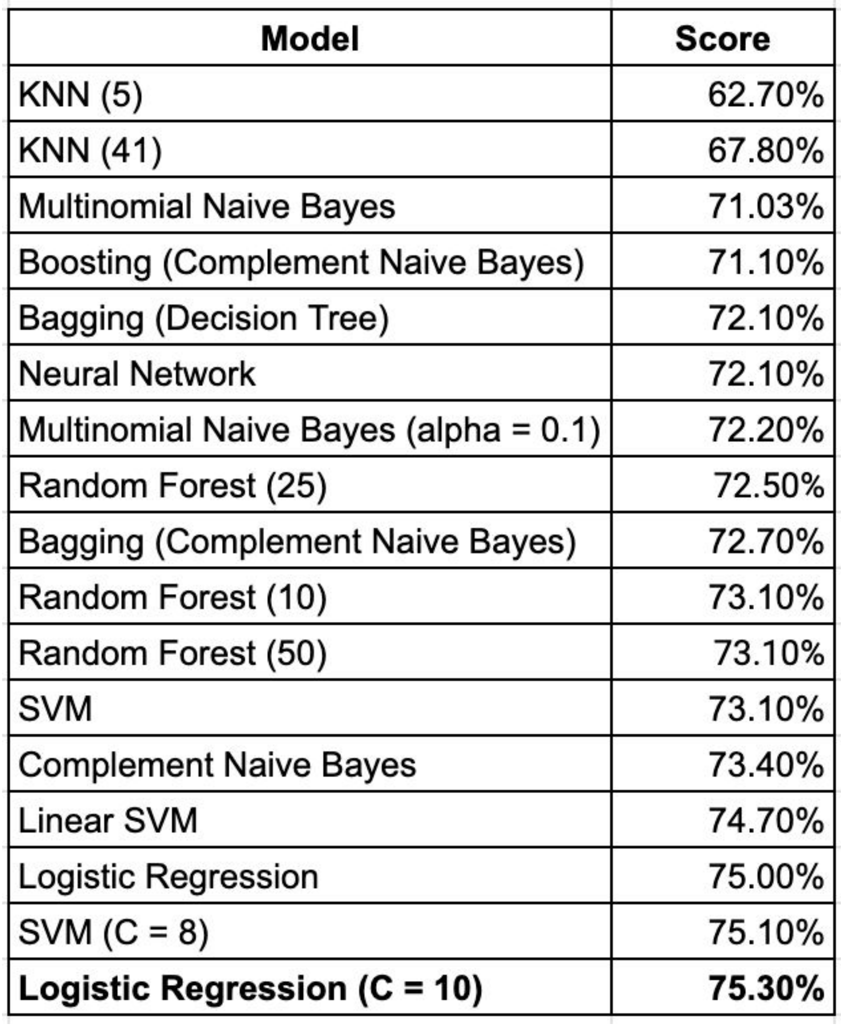
Positive Sentiment Word Cloud

These word clouds continue to show that some words (eur, company, profit) cross multiple sentiments and are high in frequency. The word clouds also show that some words are more prominent under a certain sentiment (for example, loss and 2008 in the negative word cloud).

Machine Learning Models

In order to find the best model, several algorithms were tested. The list of algorithms that were used in model building and evaluation include Multinomial Naive Bayes, Complement Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbors, and Neural Network (MLP Classifier). The Grid Search tool from Scikit Learn is being used to fine tune the algorithms to find the parameters that are best suited for training the model.

Below is a list of the different algorithms tested and the accuracy scores. The notes in the parentheses indicate what customizations were used as a result of running the Grid Search tool with the algorithm for regularization and best results. For example, the first KNN algorithm in the list used K as a value of 5 and the second used a value of 41 based on the results of the Grid Search tool. The modification of the parameter resulted in a 5% increase in accuracy.



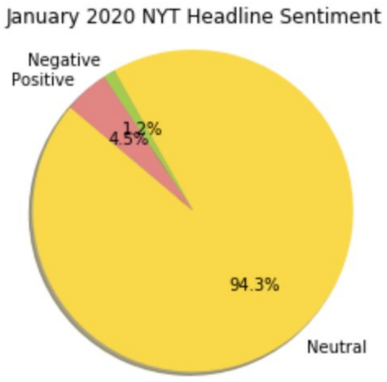
Model	Score
KNN (5)	62.70%
KNN (41)	67.80%
Multinomial Naive Bayes	71.03%
Boosting (Complement Naive Bayes)	71.10%
Bagging (Decision Tree)	72.10%
Neural Network	72.10%
Multinomial Naive Bayes (alpha = 0.1)	72.20%
Random Forest (25)	72.50%
Bagging (Complement Naive Bayes)	72.70%
Random Forest (10)	73.10%
Random Forest (50)	73.10%
SVM	73.10%
Complement Naive Bayes	73.40%
Linear SVM	74.70%
Logistic Regression	75.00%
SVM (C = 8)	75.10%
Logistic Regression (C = 10)	**75.30%**

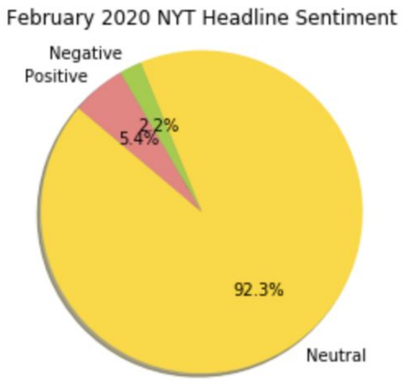
The model with the highest accuracy score was Logistic Regression with the parameter of C set to 10 (the default is 1). While 75.30% is not an extremely high accuracy score, it is a strong enough indicator to use the model to analyze the New York Times headline data.

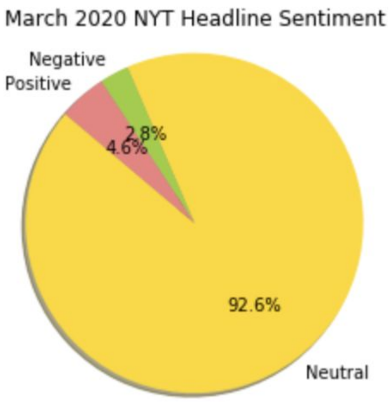
Sentiment Analysis

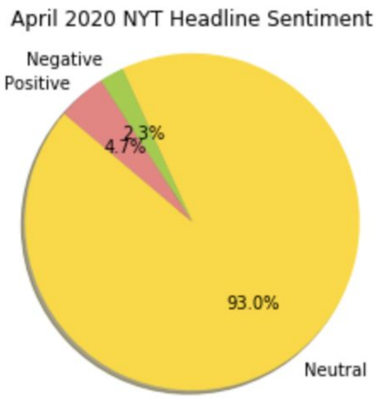
In understanding the results of sentiment labels on New York Times headlines, it is important to revisit the context of the economy for the first four months of 2020. At the beginning of the year, the United States economy was considered to be in a strong position. However, with the spread of the COVID-19 pandemic, unemployment increased dramatically in the month of March and through April. Therefore, the first four months of 2020 become an interesting capsule to do sentiment analysis on Financial News headlines and to see whether they evolved through the sudden and drastic economic changes.

For January, February, March, and April, the percentage of neutral, negative and positive headlines remained consistent. Neutral labels were assigned to the majority of the headlines (ranging from 92.3% to 94.3%) while positive and negative labels were assigned to less than 10% combined. The pie charts below show the proportion of headlines and their labels for each of the first four months of 2020.

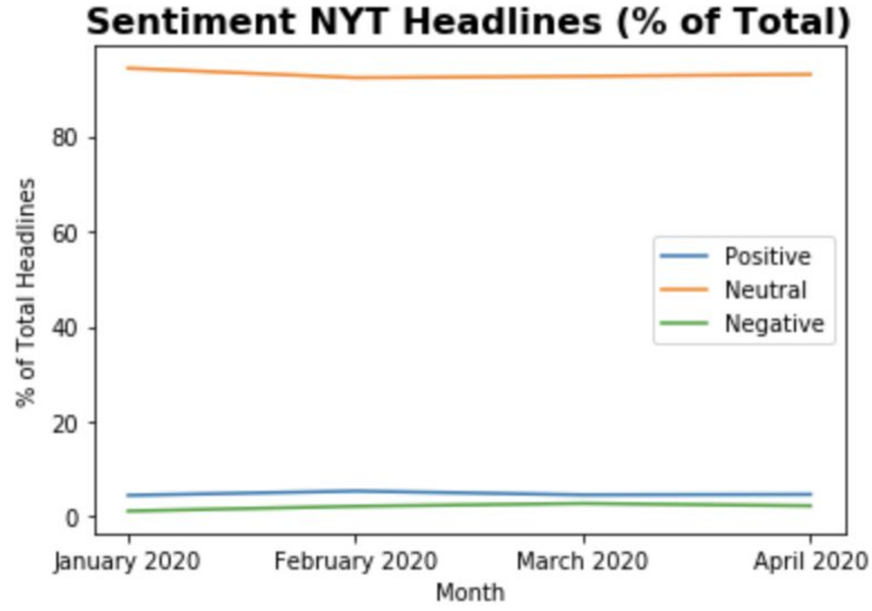




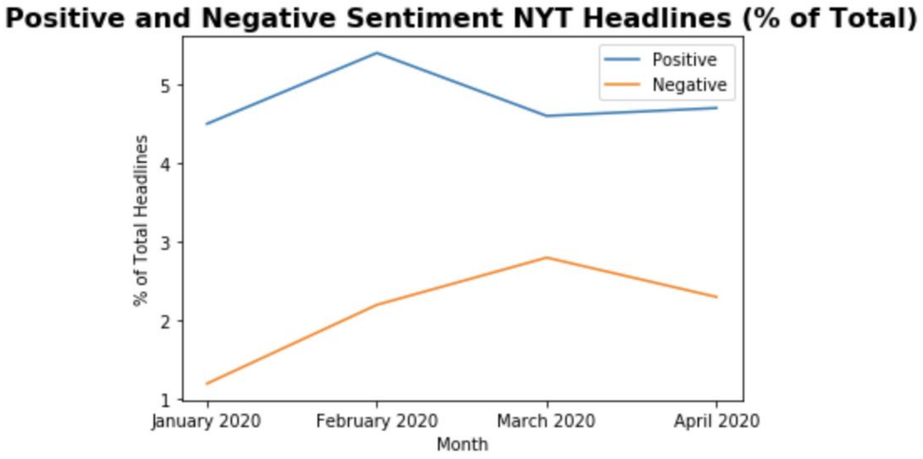




To further look at this, the below visual displays how the ratios of each sentiment changed from month to month. It is clear that the sentiments remain consistent even as there are drastic economic changes.



Upon closer inspection of the trends of Negative and Positive headlines (image below), it may appear that there are significant changes in the ratio of these headlines despite the fact that they do not follow the economic changes. However, it is important to recall that the model used has a 75.30% accuracy. It is reasonable to assume that these changes are simply due to potentially mislabeled headlines and that these sentiments continue to remain consistent.



It is important to then ask why the model consistently saw a majority of neutral headlines despite the sudden changes in the economic climate. There are two important aspects to consider here. First, the training and testing dataset held a majority of neutral headlines. It is possible that the model learned a bias of neutral headlines and is more likely to label a headline as such. Second, the neutral sentiment could be a standard of headline writing at the New York Times. Perhaps the media outlet is set on removing sentiment from the headlines of financial news articles and, therefore, the model is acting correctly on these headlines.

Further Opportunities

A project such as this can lead to some of the answers leading to more questions. If a model like the one that has been built is deployed, there are some important questions left to consider looking into when understanding how financial headlines can impact the economy or the sentiment of society.

First, it would be important to explore other media and media outlets. While the New York Times may aim for neutral headlines, twitter accounts, blogs, and other financial resources may aim to put the sentiment in the headline in order to capture the reader’s attention. It would be important to explore if social media, television media, and news media (online and print) see a similar trend in ratios of their headlines.

A second exploration ties into the first with exploring whether the neutral headline is the standard for financial news and, if so, whether this is because of its potential impact on the economy. The training and testing dataset contained a majority of neutral headlines and the subsequent model also found a majority of neutral headlines. It would be important to consider whether the data shows this is the norm for the industry and, if not, where the headlines no longer become a majority neutral.

Third, it is important to continue to build a model with a higher score of accuracy. One of the issues with this is in natural language processing. Language is filled with nuances and complications and is always changing. There is a very human element to how language plays a role in society which makes it a difficult task to teach a machine how to understand language and its role in society and humanity. With that consideration, a 75% accuracy is not a failure. However, it is not a 99% accuracy where a higher comfort level with the analysis can be provided. More headlines would need to be included in testing and training and further adaptations of the model would have to be explored.

References

[1] Baker, J. (2019 Jan 31). “Machine Learning Versus The News.” Retrieved from https://towardsdatascience.com/machine-learning-versus-the-news-3b5b479d8e6a

An article that explores how Natural Language Processing was used to analyze news stories and identify duplicate articles.

[2] Bambrick, N. (2016 Nov 18). “Using NLP and Text Mining to understand how media coverage influenced the US presidential election.” Retrieved from https://blog.aylien.com/using-nlp-and-text-mining-to-understand-how-media-coverage-influenced-the-us-presidential-election/

An article exploring how media headlines may influence society sentiment in regards to a news event (2016 US Presidential election) and how natural language processing and text mining is an important resource in completing this project.

[3] Cornelisse, D. (2020 Mar 14). “Data Science: Quantifying Stock Sentiment using Natural Language Processing on News Headlines.” Retrieved from https://medium.com/@devoncornelisse/data-science-quantifying-stock-sentiment-using-natural-language-processing-on-news-headlines-2b80f99efadd

An article discussing Data Science’s role in using Natural Language Processing to understand stock markets and the sentiment around news headlines.

[4] Czakon, J. (2020 Mar 19). “Exploratory Data Analysis for Natural Language Processing.” Retrieved from https://towardsdatascience.com/exploratory-data-analysis-for-natural-language-processing-ff0046ab3571

An article walking through how Exploratory Data Analysis strategies can be used when Natural Language Processing is applied to the dataset.

[5] Gupta, S. (2018 Jan 7). “Sentiment Analysis: Concept, Analysis and Applications.” Retrieved from https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

An article that discusses how to use Sentiment Analysis on social media data and why it is important.

[6] Martin, B. and Koufos, N. (2018). “Sentiment Analysis on Reddit News Headlines with Python’s Natural Language Toolkit (NLTK).” Retrieved from https://www.learndatasci.com/tutorials/sentiment-analysis-reddit-headlines-pythons-nltk/

A walkthrough of Python's NLTK and how to apply it to Reddit News Headlines. This will be an important resources in assessing strategies for working with the training dataset and model building.

[7] Montalenti, A. (2019 Sep 24). "Machine learning for news: the NLP engine behind Parse.ly Currents." Retrieved from https://blog.parse.ly/post/7790/machine-learning-nlp-parse-ly-currents/

An overview of a product that uses machine learning and natural language processing to understand news headlines and their influence on sentiment.

[8] Saravanou A., Stefanoni G., Meij E. (2020) Identifying Notable News Stories. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12036. Springer, Cham

A research paper exploring how to analyze a constant stream of news stories and build a machine learning model that can rank the articles by importance and look for duplicates.

[9] Shuhidan S.M., Hamidi S.R., Kazemian S., Shuhidan S.M., Ismail M.A. (2018) Sentiment Analysis for Financial News Headlines using Machine Learning Algorithm. In: Lokman A., Yamanaka T., Lévy P., Chen K., Koyama S. (eds) Proceedings of the 7th International Conference on Kansei Engineering and Emotion Research 2018. KEER 2018. Advances in Intelligent Systems and Computing, vol 739. Springer, Singapore

A research article that explores how sentiment analysis was used to analyze financial news headlines in Malaysia.

[10] Waldron, M. (2016 Jun 9). "Analyzing the structure and effectiveness of news headlines using NLP." Retrieved from https://www.datasciencecentral.com/profiles/blogs/analyzing-the-structure-and-effectiveness-of-news-headlines-using

An article that reviews how natural language processing can be used to break down how headlines are built and how then compares headlines for two journalists.