Sentiment Analysis of News Article Headlines for Stock Market Prediction

In this presentation, we'll explore how sentiment analysis of news article headlines can be used to predict the Stock Market.

Navin Sanjay 1 May 2023

Introduction

- Determine whether the Dow Jones Industrial Avg will move up or down based on News Headline feedback.

Introduction

- Determine whether the Dow Jones Industrial Avg will move up or down based on News Headline feedback.
- **Purpose**:
  - Allows for additional input over the traditional numeric comparisons.
  - The news gives a good indicator of the latest things that may affect the stock markets overall direction

Introduction

- Determine whether the Dow Jones Industrial Avg will move up or down based on News Headline feedback.
- **Purpose**:
  - Allows for additional input over the traditional numeric comparisons.
  - The news gives a good indicator of the latest things that may affect the stock markets overall direction
- **Business Case**:
  - Investors ranging from Investment firms to individual investors

Dataset

What is Dow Jones Industrial Avg?

- A stock market index that represents 30 publicly traded companies in the US
- It is used to gauge the performance of the overall stock market

Dataset

News Article Dataset

Historical Headlines were taken from Reddit (/r/worldnews).

Top 25 headlines were taken for a specific day (2008 – 2016). Ranked by users' votes

Top most headline for a given day was looked at for this project.

Stock Market Data

Dow Jones Industrial Avg (DJIA)

Target Variable:

Label 0: Stock Shifted down in price for a given day

Label 1: Stock shifted up in price/stayed the same for a given day

Dataset: Raw Dataset

	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6
0	2008-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b'150 Russian tanks have entered South Ossetia...
1	2008-08-11	1	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict'	b"Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked"'	b'What were the Mossad with fraudulent New Zea...
2	2008-08-12	0	b'Remember that adorable 9-year-old who sang a...	b"Russia 'ends Georgia operation"'	b"'If we had no sexual harassment we would hav...	b"Al-Qa'eda is losing support in Iraq because ...	b'Ceasefire in Georgia: Putin Outmaneuvers the...	b'Why Microsoft and Intel tried to kill the XO...
3	2008-08-13	0	b' U.S. refuses Israel weapons to attack Iran:...	b"When the president ordered to attack Tskhinv...	b' Israel clears troops who killed Reuters cam...	b'Britain\'s policy of being tough on drugs is...	b'Body of 14 year old found in trunk; Latest (...	b'China has moved 10 *million* quake survivors ...

Dataset: Raw Dataset

- **1989 rows x 27 cols**
- **Labelled: Minority is 46.48% (1: 1067 0: 924)**

Dataset: Processed

- Lowercase
- Remove Punctuation
- Sentence Tokenization
- Remove Stop Words
- Stemming
- Lemmatization

Dataset: Processed

Date	Label ***TARGET VARIABLE***	Top1	short ***INPUT VARIABLE***
2008-08-08	0	georgia downs two russian warplanes as countri...	georgia down russian warplane country brink war
2008-08-11	1	why wont america and nato help us if they wont...	will not america nato help will not help help ...
2008-08-12	0	remember that adorable 9yearold who sang at th...	remember adorable 9yearold sing open ceremony ...
2008-08-13	0	us refuses israel weapons to attack iran report	refuse israel weapon attack iran report
2008-08-14	1	all the experts admit that we should legalise ...	expert admit legalise drug

Exploratory Data Analysis

Sentiment Analysis

Calculating sentiment scores for each headline using natural language processing techniques.

1

2

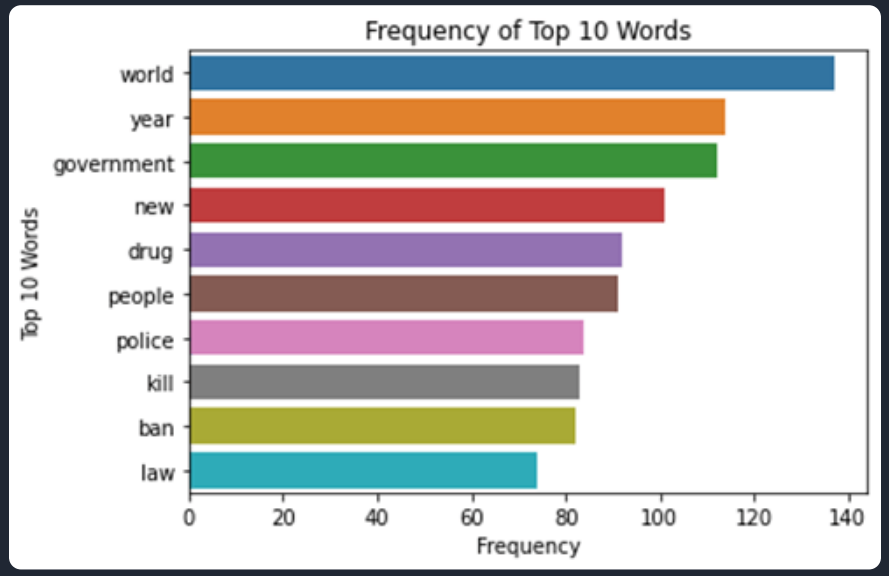
Word Clouds

Visualizing the most common words in positive and negative headlines.

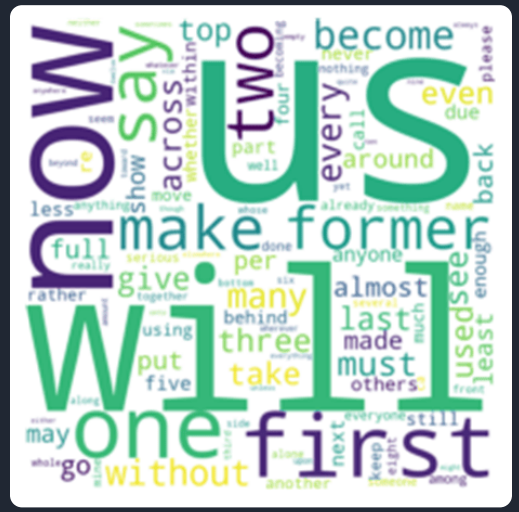
EDA: Most Commonly Used Headline Words

- The most common words seen in the dataset are expected.
  - Things such as government, new, ban, law may indicate that a change is being made which is likely to result in the stock market being affected

EDA: Most Commonly Used Headline Words



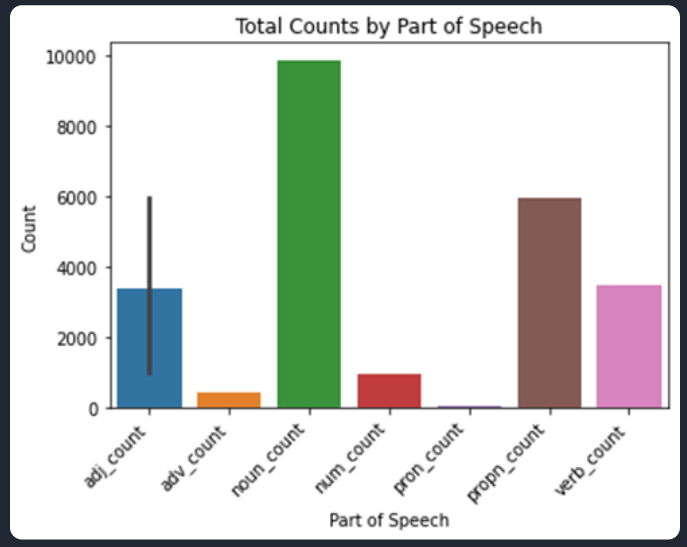
EDA: Most Common Stop Words in Dataset



- The most common stop words of the dataset was looked at to see if any need to be inputted back into the dataset
    - *None were considered of importance (No stop words were revoked)*

EDA: Popular Part of Speech Tags

- There is a significant amount of nouns (Common and Proper) in headlines of news articles. Makes sense as articles are generally about *something/Someone*.
- The nouns will play an important role in sentiment analysis, as if something related to a stock gets mentioned in the headline it will have a large impact on how people see that stock



Feature Engineering

Vectorize

Converting text data into numerical representations (vectors) that can be processed by machine learning models.

NLP Features

Extracting features such as named entities and part-of-speech tags.

Feature Engineering

Vectorization Techniques Used

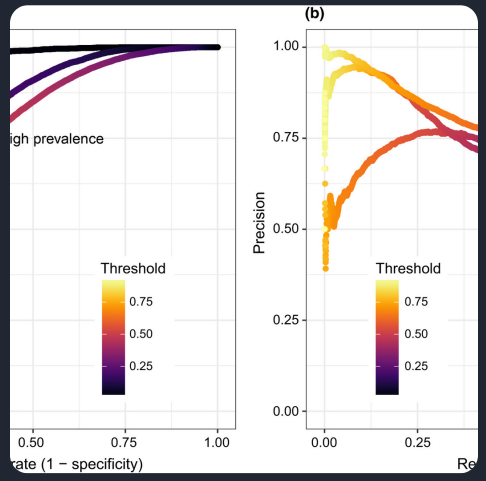
- Count Vectors
  - *Representing a document by counting the frequency of each word in it.*
- TF-IDF Vectors
  - *Assigning weights to words based on how frequently they occur in a document and how **common** they are across all documents in a corpus.*
    - Word Level
      - *Individual words as features.*
    - Ngram Level
      - *Adjacent sequences of n words as features. For example, bigrams (n=2) or trigrams (n=3).*
    - Characters Level
      - *Individual characters as features. This can be useful for detecting patterns in text like misspellings or emoticons.*

Models

Classification Problem

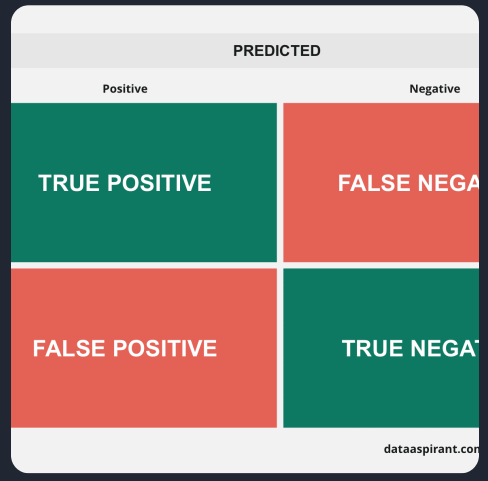
1. Logistic Regression
2. Naive Bayes
3. Support Vector Machine
4. Random Forest Classifier
5. Gradient Boosting Classifier

Results and Evaluation



Accuracy Score

Evaluating model performance using the prediction Accuracy Score.



Confusion Matrix

Evaluating model performance using a confusion matrix

Results: Accuracy Score

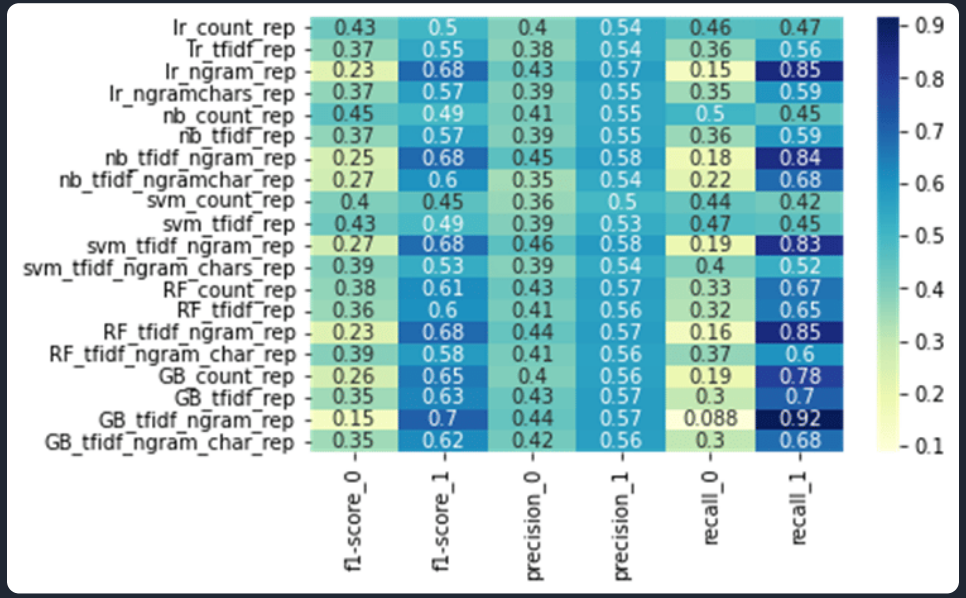
	Count Vectors	WordLevel TF-IDF	N-Gram Vectors	CharLevel Vectors
LogisticRegression	0.46	0.47	0.55	0.49
Naïve Bayes	0.47	0.49	0.56	0.48
SVM	0.43	0.46	0.56	0.47
RF	0.51	0.53	0.56	0.49
Boost	0.53	0.53	0.57	0.52

Results: Accuracy Score

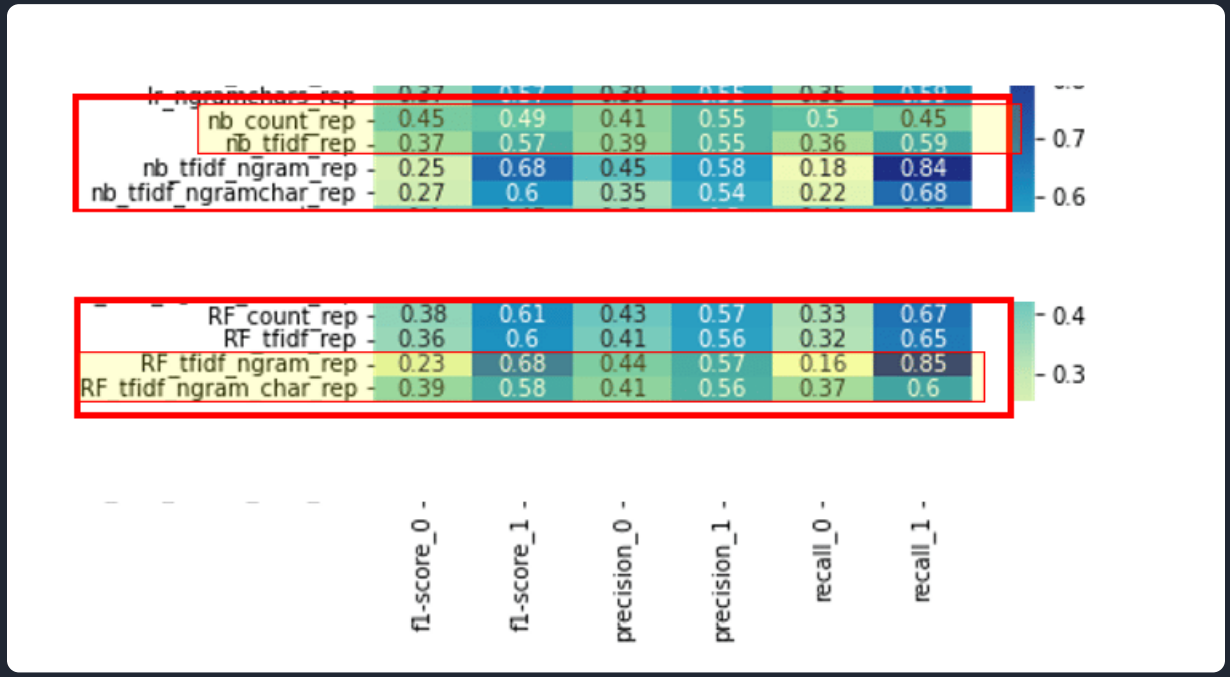
	N-Gram Vectors
LogisticRegression	0.55
Naïve Bayes	0.56
SVM	0.56
RF	0.56
Boost	0.57

Results: Confusion Matrix Summary

Heatmap displaying how well the models performed at prediction
True Positives, True Negatives etc

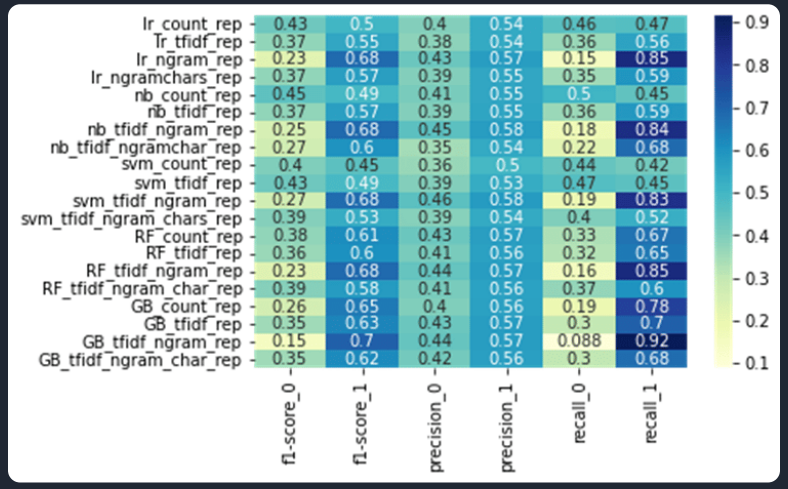


Results: Confusion Matrix Summary



Results: Confusion Matrix Summary

- Model performed better at predicting Label 1 (Stock shifted up in price or Stayed the same) than for Label 0 (Stock price shifted down)
- The models are better at predicting when the stock will go up or stay the same, rather than predicting when it will go down.



Results: Confusion Matrix Summary

- Best Model based on Weighted-avg F1-Score
  - Gradient Boosting using TF-IDF Word Vectorization
  - Random Forest using Count Vectorization

	weighted-avg_f1-score
GB_tfidf_rep	0.51
RF_count_rep	0.51
svm_tfidf_ngram_rep	0.50
GB_tfidf_ngram_char_rep	0.50
RF_tfidf_rep	0.49

Results - GridSearchCV

	N-Gram Vectors
NB	0.56
SVM	0.57
RF	0.56
Boost	0.57

Summary

Objective:

Predict whether the Dow Jones (DJIA) will move up or down based on sentiment analysis of news article headlines.

- Text classification using NLP was performed on news article headlines from Reddits /r/worldnews between 2008-2016. The top headline was taken for each day. DJIA data was used as the target variable, with labels of 0 for a downward shift in price and 1 for an upward shift/staying the same.
- The classifications models used included: Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, and Boosting. Grid search cross-validation was performed to find the best hyperparameters for each model.

Summary

1 Main Findings

Despite the difficulty of predicting the stock market, our models showed promising results using sentiment analysis of news headlines.

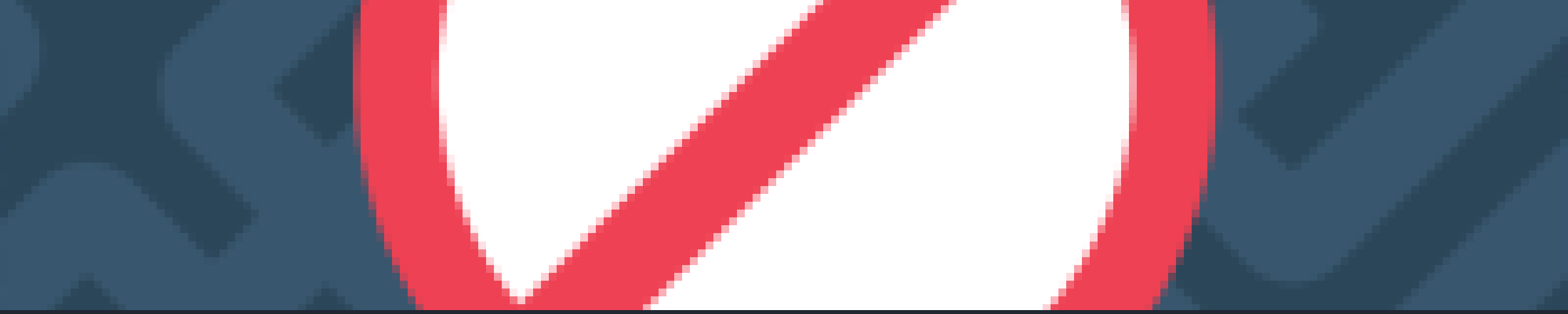
The best performing models were:

- Naive Bayes, Support Vector Machine, Random Forest and Gradient Boosting

Models performed very similarly so more work must be conducted to get a conclusive statement of which model is the best to use.

2 Future Work

Further research could be conducted on incorporating additional features such as social media sentiment and economic indicators to improve model performance.



Conclusion

News article headlines can be used as a supplement to traditional numerical analysis for predicting the movement of the stock market. While the accuracy of the models is not high enough to rely on solely, it can provide additional insights for investors to make more informed decisions.