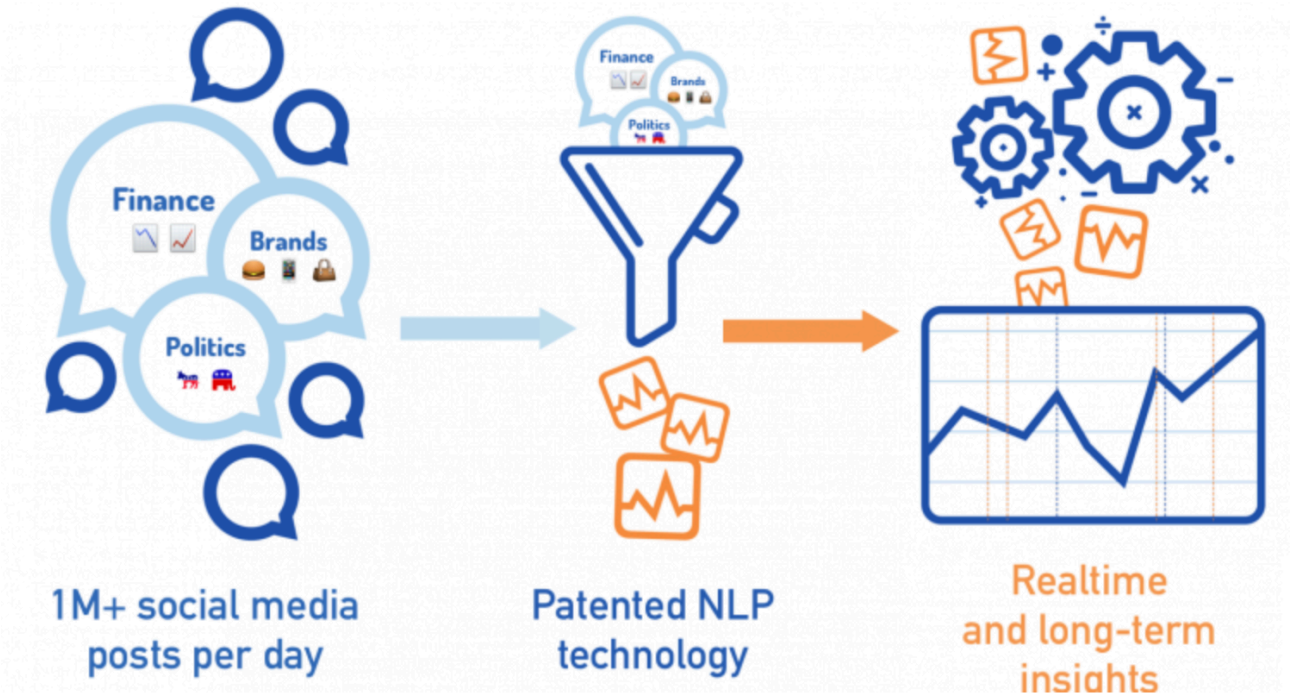


Minimum Viable Product

Prepared for: Metis Staff Instructors(Chris Bruehl, Leon Johnson)
Prepared by: Rahul Raju, Metis Student
September 14, 2021

EXECUTIVE SUMMARY

Objective

An initial proposal was made to aid WallStreet firms in incorporating textual information into existing trading algorithms via Natural Language Processing. The project has since been expanded conceptually to include a topic modeling analysis as well. Topic modeling could aid research analyst in identifying what factors led to certain historical market trends.

Data Preprocessing

- Data was broken into training and test data for the supervised portion of the project
- A separate, undivided data set was maintained for the unsupervised portion of the project
- For each day the top 25 headlines, as voted on by Reddit users, were combined into one document
- Punctuation marks, numbers, stop words and capital letters were removed or altered
- Data was inputed into a Count Vectorizer which returned a document-term matrix
  - Within Count Vectorizer data was tokenized and stemmed
- Principal Component Analysis was conducted in order to reduce dimensionality

Analysis & Product

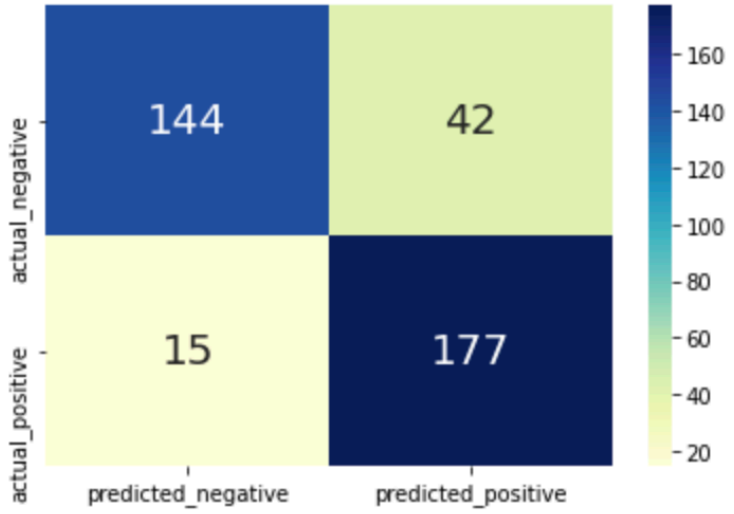
Numerous models were created, first by experimenting between CountVectorizers and TFIDF Vectorizers. Within each vectorizer, parameters were tested across multiple values. These parameters included bi-grams, binary versus actual counts and the document frequency of terms. The outputs were used in two different classifications models, *logistic regression* and *random forest*. *Accuracy* was chosen as a metric as it is possible to lose and make money regardless of the direction of the

market due to shorting and equity derivatives. Thus correctly predicting either direction is equally beneficial to traders. The table below displays the results of the various models evaluated so far.

VECTORIZER & PARAMETER	CLASSIFIER	SCORE
TFIDF Vectorizer	Random Forest	0.849
Count Vecotorizer with Uni & Bi-gram	Random Forest	0.847
Count Vectorizer, Uni & Bi-gram with PCA	Random Forest	0.847
TFIDF Vectorizer, bi-gram, with PCA	Random Forest	0.823
Count Vecotorizer	Logistic Regression	0.815

Logistic Regression was abandoned early on as *random forest* produced the optimal results. Up to this point in the process, the TFIDF Vectorizer with a random forest classifier has been the best performer with an accuracy score of 0.849. The figure below is the confusion matrix for this model as well as some supplementary metrics.

```
Accuracy: 0.849
Precision: 0.808
Recall: 0.922
F1: 0.861
```



Further Work

- More models with varying parameters will be tested
- A gridsearchCV will be performed to optimize the parameters of the classifier used on the optimal model
- The process of topic modeling will begin, with various methods explored, including; Latent Semantic Analysis, Non-Negative Matrix Factorization, and Latent Dirichlet Allocation
- A function will be created that returns topics associated with a particular day or a range of days inputed into the function