Sentiment Analysis on Finance Headlines from the New York Times

David Suffolk

Overview

- Introduction of Problem
- Project Objectives
- Data Preparation
- Data Exploration
- Model Building
- Sentiment Analysis
- Further Opportunities

Introduction of Problem

- Society’s attitude about economics can impact economy
- 2020 has seen a major shift in the economy due to COVID-19
- Can the sentiment of the United States be measured?
- Does the sentiment reflect the changes in the economy?

Project Objectives

- Use machine learning to build a sentiment analysis model
- Focus on finance headlines
- Apply model to New York Times headlines of 2020
- Data Sources
  - Sentiment Analysis for Financial News Headlines (Kaggle)
  - New York Times API

Data Preparation

- Financial News Headlines
  - Headline
    - Text
  - Sentiment
    - Negative
    - Neutral
    - Positive
    - Converted to numeric for machine learning algorithms

Data Preparation

- Headlines
  - Remove punctuation
  - Tokenize words (NLTK)
  - Remove stopwords (NLTK)
- Original: “According to Gran, the company has no plans to move all production to Russia although that is where the company is growing”
- Modified: ““‘According', 'Gran', 'company', 'plans', 'move', 'production', 'Russia', 'although', 'company', 'growing'”

Data Preparation

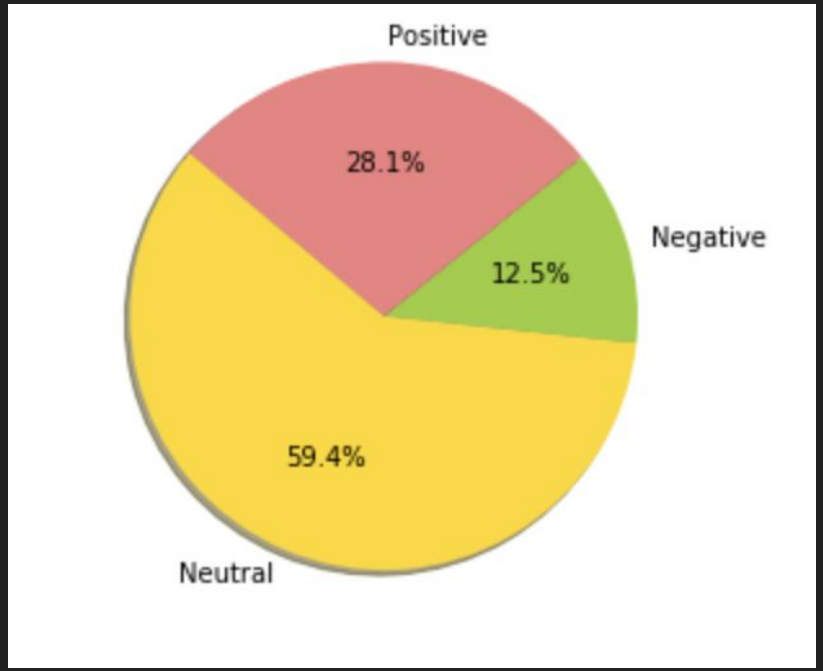
- Headlines
  - Train/Test (80/20)
  - Tfidf Vectorizer
    - Ngram range - 1,2
    - Max Frequency - 0.9
    - Min Frequency - 0
    - Max Features - 4000

Data Preparation

- New York Times API
  - Limit of 10 requests per minute and 4000 per day
  - Created datasets per month (January, February, March, April of 2020)
  - Keyword “finance”
  - Same preparation steps as training/test dataset

Data Exploration

- Sentiment

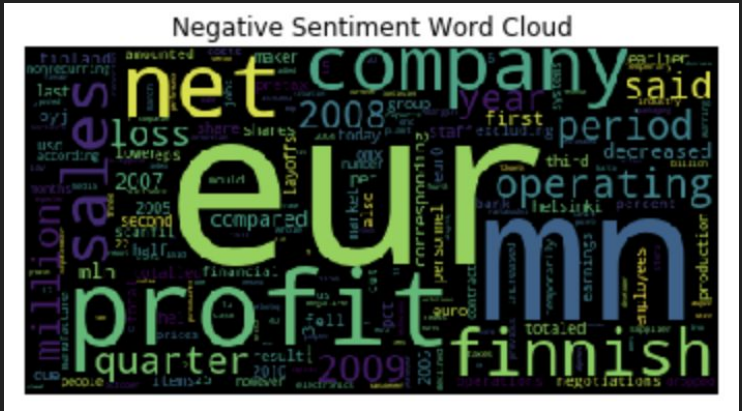


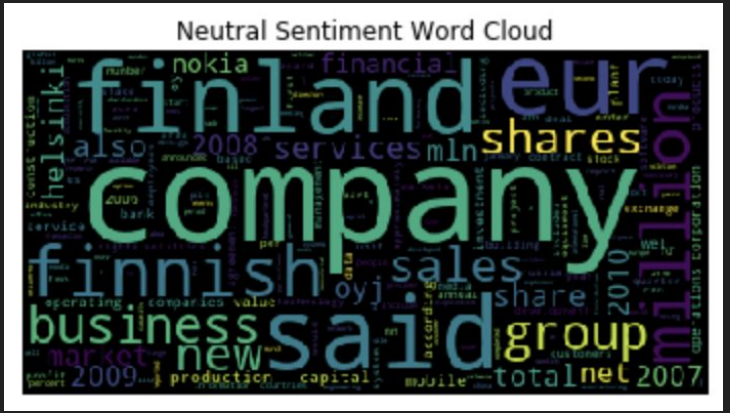
Data Exploration

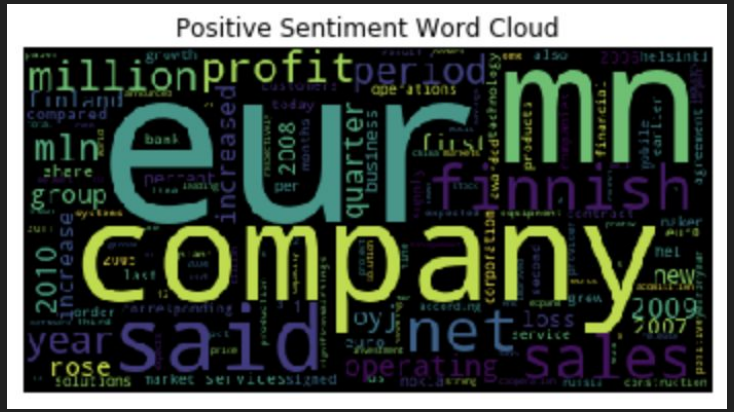
- Word Frequency (Python WordCloud Library)

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

Data Exploration







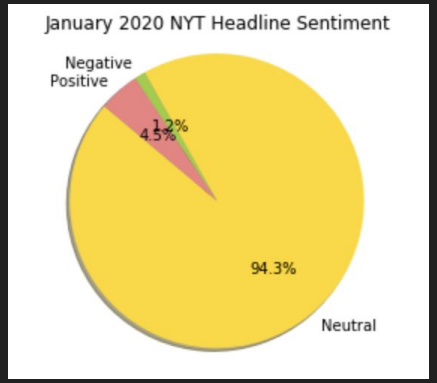
Model Building

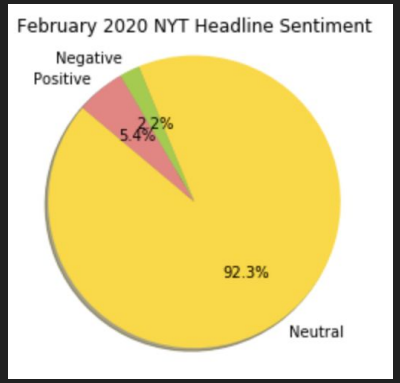
- Multiple algorithms used to find best model
- Including ensemble models
- Scores were approximately the same
- Grid Search Tool (SciKit-Learn) to compare parameters for best model and avoid overfitting

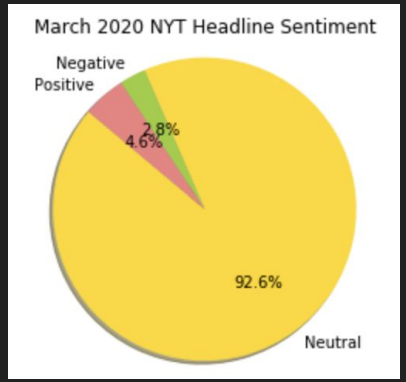
Model Building

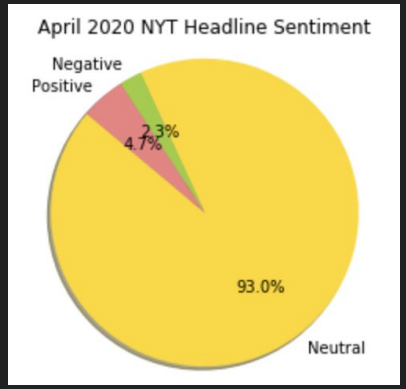
Model	Score
KNN (5)	62.70%
KNN (41)	67.80%
Multinomial Naive Bayes	71.03%
Boosting (Complement Naive Bayes)	71.10%
Bagging (Decision Tree)	72.10%
Neural Network	72.10%
Multinomial Naive Bayes (alpha = 0.1)	72.20%
Random Forest (25)	72.50%
Bagging (Complement Naive Bayes)	72.70%
Random Forest (10)	73.10%
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%**

Sentiment Analysis

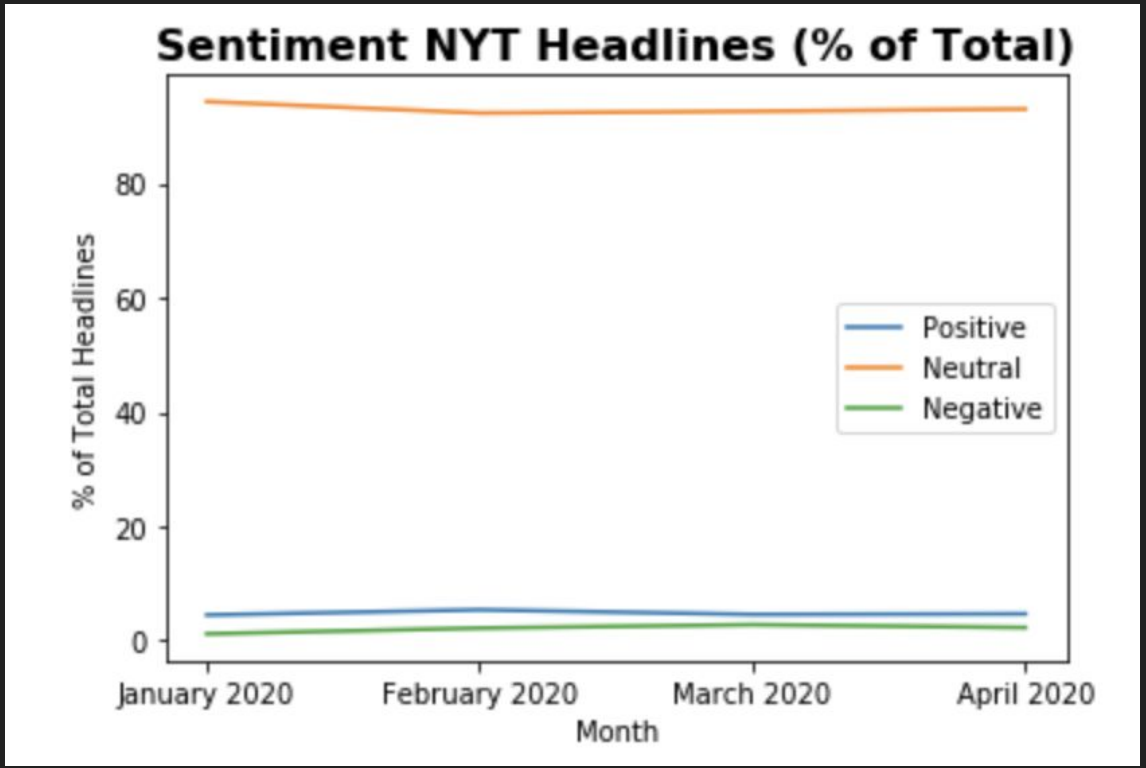




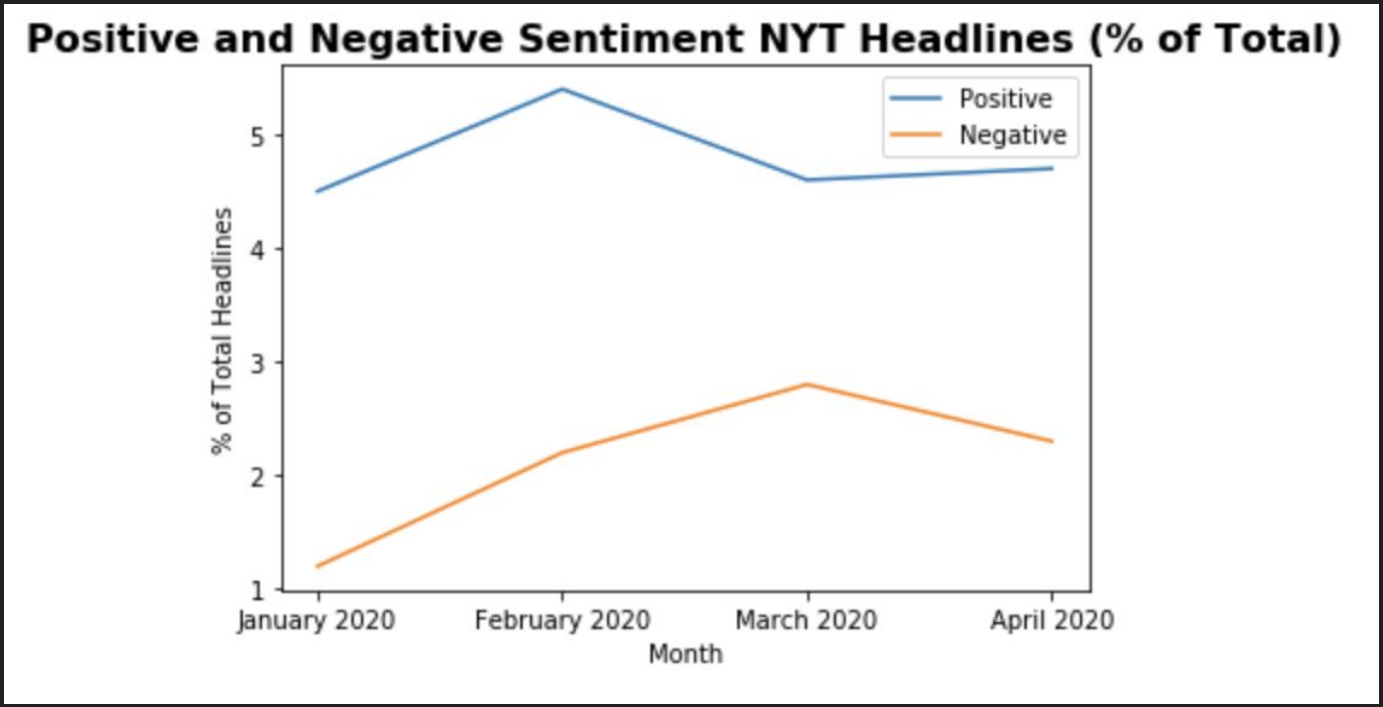




Sentiment Analysis



Sentiment Analysis



Further Opportunities

- Explore other media
- Explore if neutral is the standard across all financial news
- Continue to train model for better accuracy