Sensitive Stonks

Analyzing Stock Market Prices with Twitter Sentiment

Santiago Carrera, Helen Hua, Andrea Koltai, Maria Seoane, Stephanie Stevens

Georgia Tech

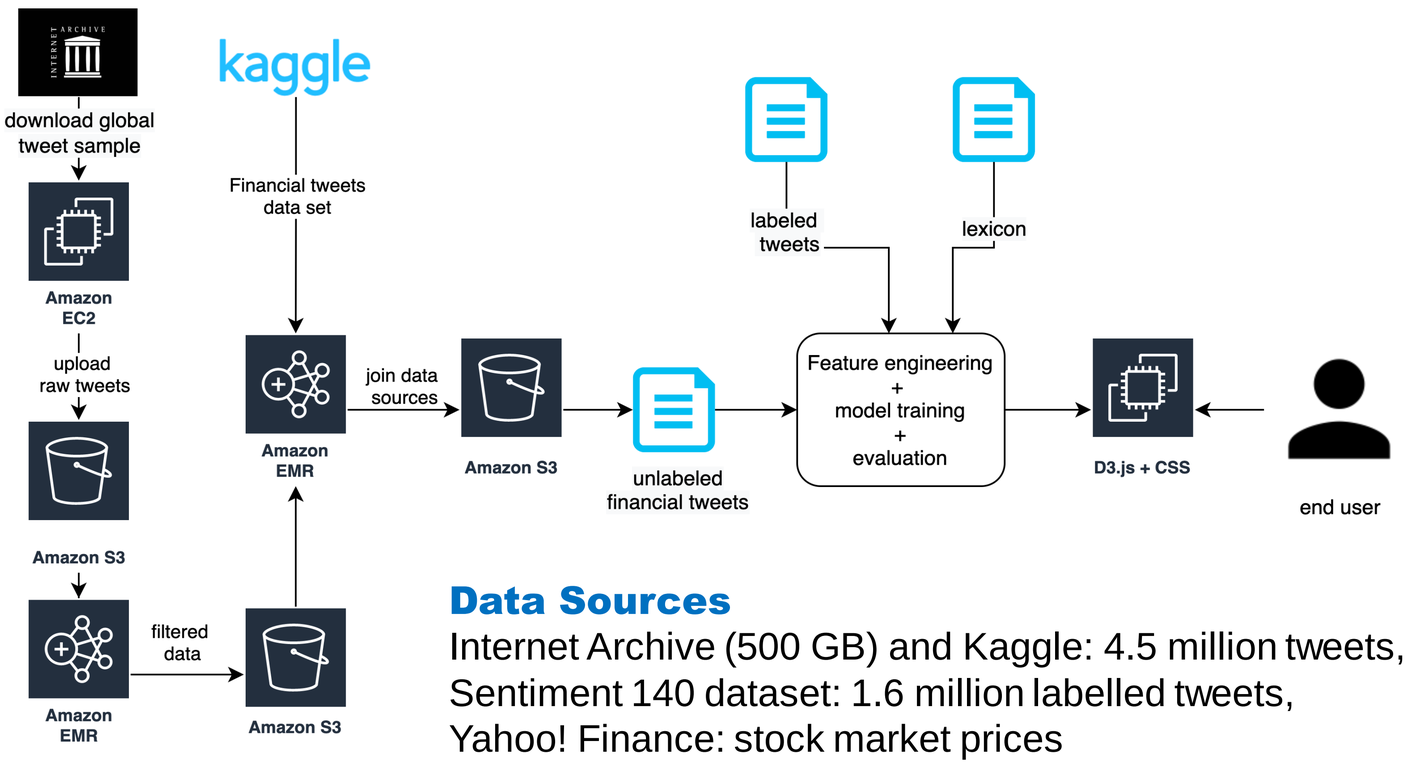
Motivation

The motivation behind this project was to build a web application that would allow a broader audience to have access to the insights that Twitter sentiment can provide on stock performance.

Summary

The objective of our project is to create a tool to illustrate the relationship between stock price and aggregated tweet sentiment through time, as well as suggest Twitter accounts to follow for insight on specific stocks. This can be a useful tool for amateur investors wishing to incorporate this information in their investments, professional investors looking to gain a competitive advantage, and companies interested in understanding how public sentiment of their company relates to their performance.

System Overview

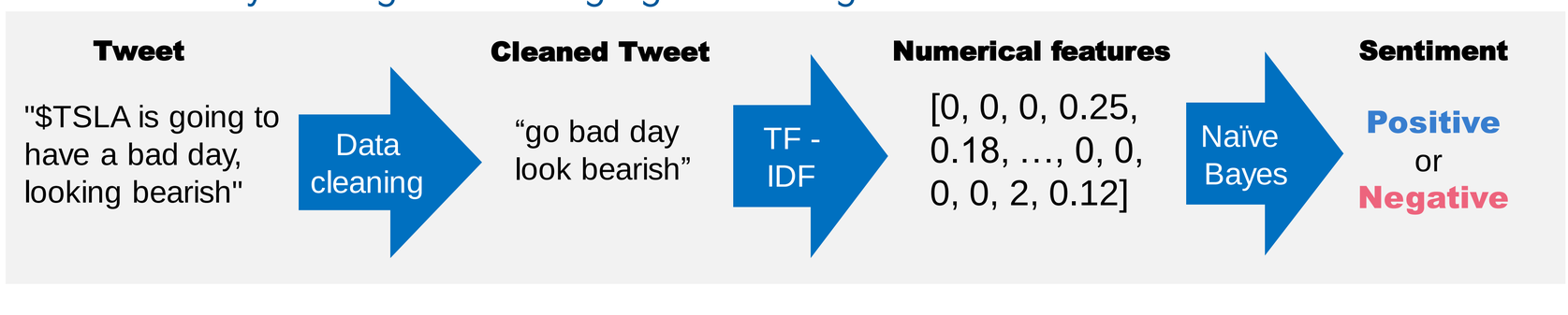


Data Sources

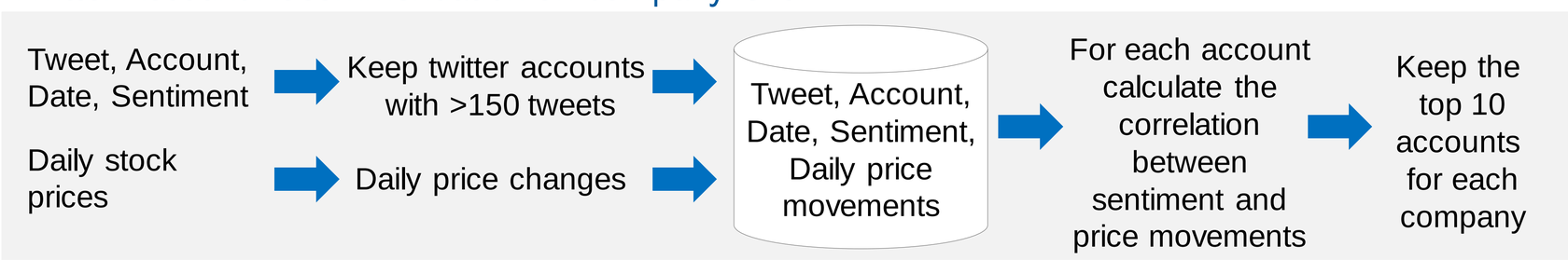
Internet Archive (500 GB) and Kaggle: 4.5 million tweets,
Sentiment 140 dataset: 1.6 million labelled tweets,
Yahoo! Finance: stock market prices

Models

Sentiment Analysis using Natural Language Processing



Twitter Account Recommendation on company level



Term Frequency - Inverse Document Frequency (TF-IDF) is a vectorization to create feature values. TF-IDF measures the importance of the words by comparing the frequency of the word that appeared in a document (term frequency) with the number of documents the word appeared in (inverse document frequency).

Naïve Bayes Classifier is based on Bayes Theorem and uses the assumption that the features are conditionally independent. This makes possible the simple calculation of probability of the output class (y) given a feature set X=(x1, x2, …):

$$P(y \mid X) = \frac{P(x_1 \mid y)P(x_2 \mid y)\ldots P(x_n \mid y)P(y)}{P(x_1)P(x_2)\ldots P(x_n)}$$

Experimentation & Evaluation

We tested 3 different models to predict tweet sentiments: Naïve Bayes, Support Vector Machine, and Neural Networks. We decided to use Naïve Bayes model due to its more efficient runtime, lower model overfitting and still very good predictive power.

Model	Train accuracy (%)	Test accuracy (%)	Runtimes (s)
Naïve Bayes Classifier	79.65	73.6	0.935
Support Vector Machine	86.50	73.4	233
Neural Network	92.11	70.65	235

The user experience evaluation showed that the users usually found the tool and the provided information helpful, and a fraction of users would use it to integrate this information into their trading decision. The users pointed out the lack of visible correlation between the tweets and prices. This is in line with our results that shows only moderate (0.1 - 0.25) correlation between tweet sentiment and price movements.

Visualization

The tool allows the user to select a stock symbol and time period for which it shows the daily overall Twitter engagement and its predicted sentiment: count of positive (blue) and negative (red) tweets, as well as the daily stock performance. On the right, we have included recommended Twitter accounts to follow for the selected stock symbol.

Future improvements

The model's ability to detect relevant tweets could be improved to show a clearer picture of actual Twitter engagement. The tool could also be extended to cryptocurrencies as their market is influenced by social media.

Sensitive Stonks