MGT 6203

Sentiment Analysis for Stock Price Forecast

Maria Daniela Sanchez - msanchez86
Galen Hew - ghew3
Haden Brearton - gbrearton3
Samiksha Mailarpwar - smailarpwar3

Fall 2022

TABLE OF CONTENT

01
OVERVIEW

02
METHODOLOGIES

03
MODELING

04
CONCLUSIONS

01 OVERVIEW

BACKGROUND

Sentiment analysis according to Sentiment Analysis of Twitter Data: A Survey of Techniques is the process of automating the mining of attributes such as attitudes, opinions, views, and emotions from different tweets by using NLP techniques.

Feature extraction, as the authors described is tagging certain words and their frequencies, position of the words aka within a line of text because it can determine how much that word affects the sentiment of the tweet, identifying and tagging opinion words and phrases, tagging negation terms, etc.

Also, Sentiment Analysis for Predicting Stock Price Movements article presents BERT as a deep learning technique which we actually chose to utilize in our approach.

INTRODUCTION

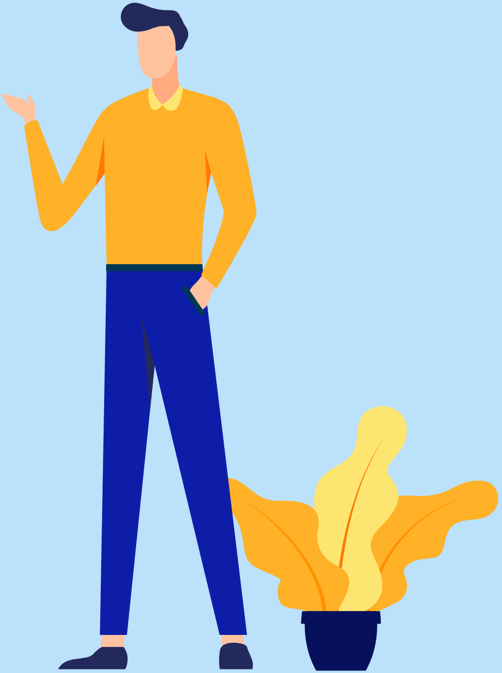
Our goal is to predict stock prices (S&P or for specific companies) based on sentiment analysis. There are many variables that can influence the price of a stock. As seen from the sentiment-driven price movements of stocks such as Gamestop and AMC, there is reason to believe sentiment should have some correlation with the stock market. Therefore we would like to identify the effect of public sentiment expressed on social media on the stock market. Due to its ease of extracting data and its reach, we will use data from twitter.

Can we forecast stock market prices based on social media sentiment analysis?

INITIAL HYPOTHESIS

We expect our model to show us that betting against crowd sentiment can forecast stock prices in a way that will enable us to outperform the market.

02 METHODOLOGIES



METHODOLOGIES

1. Pre-process the text data.
2. Use supervised models as our baseline approach.
3. Use modern approaches such as a BERT (Bidirectional Encoder Representations from Transformers) model.
4. Train the models by first splitting them into train and test data sets.
5. Run more computationally intensive k-fold cross validations on the models.
6. Compare the models

DATA

1. The dataset we ended up using is the financial phrasebank dataset available on hugging face and tensorflow. This data set consists of 4,840 sentences selected from financial news and annotated for Positive, Negative, and Neutral sentiment by 16 different annotators with experience in the financial domain.

2. The variables that we used were the thread id, dates, tweet sentence, sentiment labels.

DATA CLEANING

1. We started the data cleaning process by removing any special characters and stopwords such as articles and pronouns using NLP which enabled us to strip text down to the bare minimum to analyze it more efficiently and accurately.
2. We then encoded the labels into numbers. Most of the labels (65%) are in the neutral category, while the rest are in positive and negative labels.
3. Since BERT works with fixed-length sequences we truncated our sentences at 160 characters.
4. We also use an attention mask as a binary tensor indicating the position of the padded indices so that the model does not attend to them.

03 MODELING

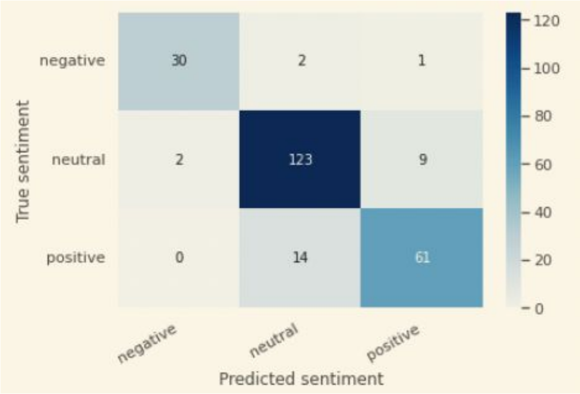
TRAIN TEST

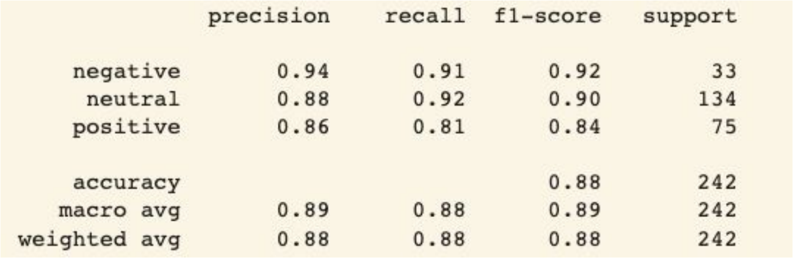
On the BERT model, we used a moderate learning rate of 2e-5. We ran through 10 epochs which took 20 min to train on a Nvidia K80 GPU. Since we will fine-tune our model, we utilize Adam as our optimizer, which outperforms stochastic gradient descent (SGD).

Having saved the trained model, we realised we required more data representative of the S&P500. Therefore we scraped comments and threads from reddit boards where investors/punters congregate and express their views freely, such as Wallstreetbets. We did this through Reddit's open source API. It takes a long time to ingest the data, thus we saved the comments as pickle file for retrieval.

EVALUATION

The confusion matrix shows us that for each category, true negative, neutral and positives are high. Neutral comments are likely to be taken as positive, perhaps because we have more positive samples than negative ones as well. The precision, recall and f1-scores are close to 90%. This could be due to overtraining as well since our dataset is less than 5000.



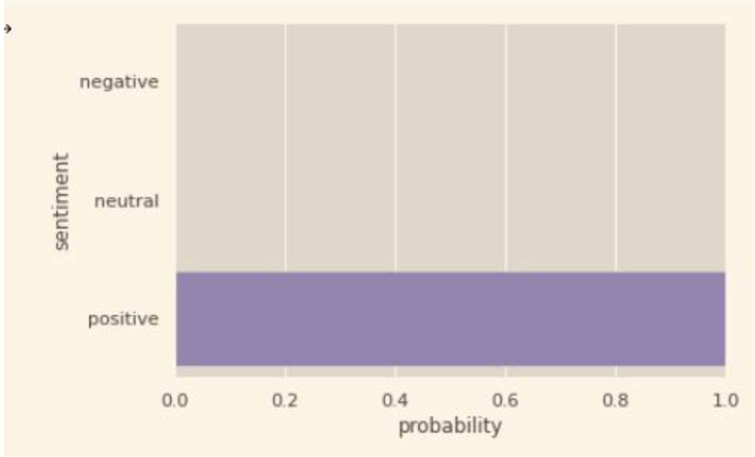


```
              precision    recall  f1-score   support

    negative       0.94      0.91      0.92        33
     neutral       0.88      0.92      0.90       134
    positive       0.86      0.81      0.84        75

    accuracy                           0.88       242
   macro avg       0.89      0.88      0.89       242
weighted avg       0.88      0.88      0.88       242
```

EVALUATION

(ADP News) - Nov 5 , 2008 - Finnish electronic measurement products and solutions maker Vaisala Oyj (OMX : VAIAS) said today that its net profit rose to EUR 18 million (USD 23.1 m) for the first nine months of 2008 from EUR 1

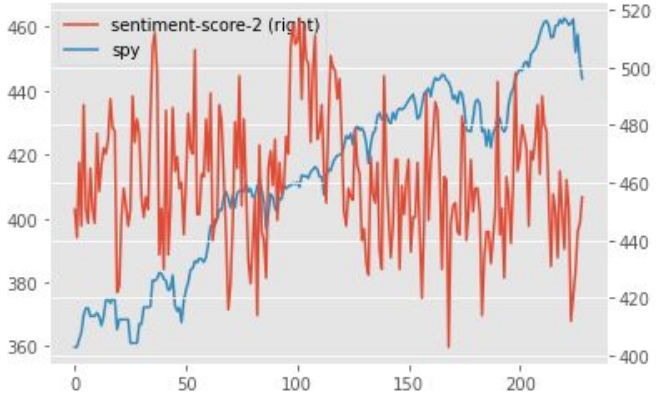
True sentiment: positive



APPLICATION ON SPY

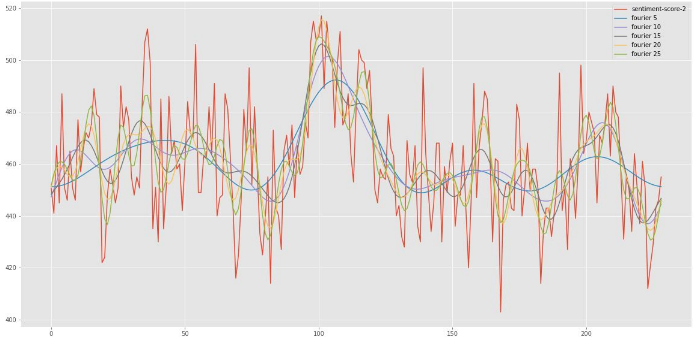
01

With sentiment output from our model, we want to see whether sentiment has a relationship with the S&P500 thus we plot them against each other . We obtain the S&P500 data from the financial functions for python library (ffn). However we can see here that it is difficult to interpret any pattern at all compared with the SPY index.



02

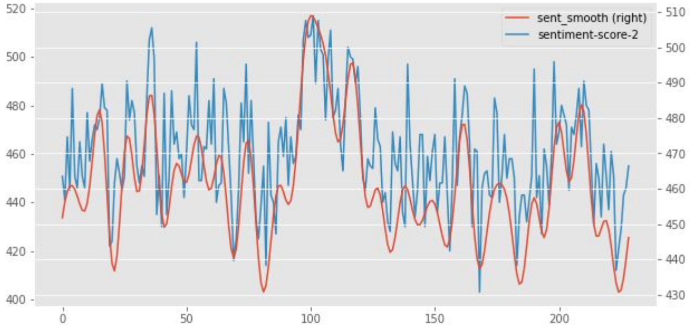
It turns out that there is too much noise and we cannot visually see any relationship. Therefore we turn to fourier transform decomposition to remove the noise. We look at fourier transformations by steps of 5 days, from 5 to 30 days.



APPLICATION ON SPY

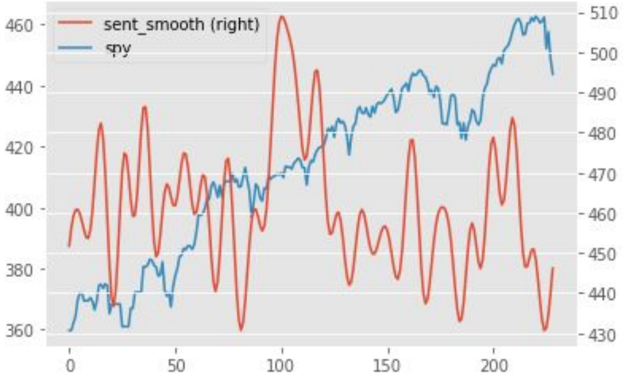
03

From the fourier at 25 days, we are able to see that sentiment trends with the market and might be a leading indicator of the market, assuming little leakage effect from the transformation. We then convert the fourier output into real numbers for ease of use, labelled as sent-smooth in the diagram below.



04

As we can see from here, a strategy could be to counter trade the sentiment at extreme peaks and troughs, assuming that sentiment precedes the price.



04 CONCLUSIONS

IMPROVEMENTS

To improve the project, we should look into the aggregation method for the sentiment of each comment. Right now we simply take the mean for each day. However there are many bots that post meaningless comments and many to manipulate or sway public opinion. Whether these work in effect to their cause or should be removed is something we can explore. We can also look at weighting the comments based on number of stars/ followers and retweets.

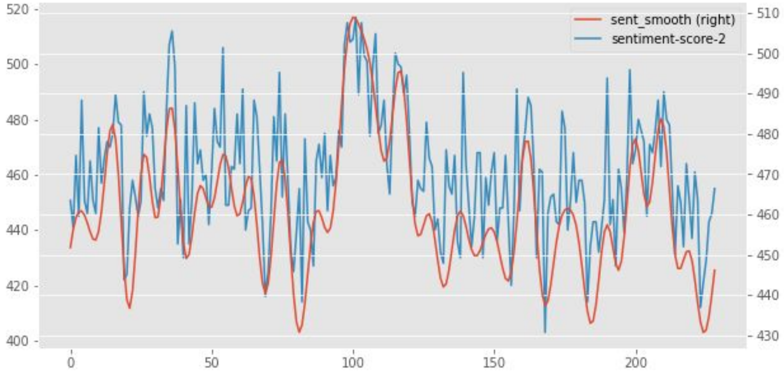
Further, we should use at least 7 years of stock market data to backtest, to incorporate market cycles and black swan events. Due to time limitations, we are only able to obtain a year of financial data which is scarcely sufficient to make statistical conclusions. We did not backtest the strategy as it seems out-of-scope and backtesting is not statistically significant either.

CHALLENGES

We have chosen models necessary to craft the solution are a bit out of the scope of what we have learnt in the class. Nevertheless, it is a challenge we have taken up and it has been definitely interesting and informative to learn new models such as the data cleansing techniques such as tokenization, lemmatization, and removal of stopwords in addition to BERT, SHAP and LSTM.

KEY FINDINGS

The graph below is the output and here, we can clearly see that the sentiment line (blue) trends very closely with the market (red). This indicates that as predicted in our initial hypothesis, sentiment is a very good, in fact a strong, indicator of the market. This is exactly in line with our literature surveys and understanding of social media's influence on predicting short-term stocks.



CONCLUSIONS

- Social media sentiments are hugely tracked and analyzed by hedge funds and large investment bankers as it does indeed have some correlation with the market
- Causation cannot be proven, but as our analysis shows, Twitter sentiments do have a strong correlation with how the market behaves and can be argued that it may be one of the leading indicators of market performance.
- This particular area is extremely promising and with advancements in natural language processing, sentiment analysis can most definitely get better at predicting how the stock market will sway.

SOURCES

[1] Vishal A. Kharde, S.S. Sonawane. Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications, vol. 139, no. 11, April 2016.

[2] Kevin Hu, Daniella Grimberg, Eziz Durdyev. Twitter Sentiment Analysis for Predicting Stock Price Movements, Department of Computer Science, Stanford University.