Abstract

The projects goal is to incorporate sentiment analysis result of tweets related to specific companies into the prediction for the companys performance, i.e. their stock prices, and see how much weight user feedbacks on twitter have in the model. For sentiment analysis task, we compared Valence Aware Dictionary and sEntiment Reasoner (VADER) with Naive Bayes model and chose the better one to generate intermediate sentiment scores. For stock prediction, we implemented a Long short-term memory (LSTM) model, and compared models taking into different predictors, i.e. tweets count and sentiment score.

1. Introduction

The focus of the project is to analyze whether company-related tweets could improve the prediction power of a stock prediction model. The project is divided into two stages: the first is to gather relative data and to get an intermediate score by performing sentiment analysis, and the second is to apply the score in the potential prediction model.

Although extensive amount of resource has been dedicated to the sentiment analysis of long paragraphs of text, performing sentiment analysis on text with a relatively short body remains challenging. Compared with other sources of text used in sentiment analysis such as IMDB movie review (Pang & Lee, 2004), tweets have their own advantage and disadvantages. The informal nature of tweets means that sometimes the part of one given sentence where most of the sentiment lies is a special character sequence that represents a specific emotion (such as D: or :) ) or hashtags (such as feelsbadman or monkaS). This means lexicon based models that focus mainly on formal expressions may not be the best bet. With a maximum length of 140 characters, most of the tweets contain only one or two sentences, thus making models such as ANN and LSTM less efficient. So in earlier researches, emoticons and hashtag are either not included in the sentiment analysis or another dataset was used to index them (Kouloumpis et al., 2011). After careful consideration, the Naive Bayes model was chosen as the model for sentiment analysis in this project, being that it is very easy to train and the ability to incorporate the sentiment value of emoticons and hashtags into evaluation.

The research about combining sentiment analysis to predict stock price has been an interesting topic for both academics and financial practitioners. Previous research involving using Neural Network Autoregressive model (NNAR) (eff, 2018) and linear regression model (Cakra & Trisedya, 2015) to predict stock price with and without sentiment score. While some research indicates that using news and tweet sentiment score from Bloomberg and combine with NNAR model can outperform 65% of NNAR model without sentiment score. Since recent research indicates that LSTM model can be efficient in predicting time series compared to ARIMA or linear regression (Singh, 2018), we wanted to test if adding sentiment score to LSTM model will also improve the prediction power as in linear models.

2. Method

We used a large dataset Sentiment140 (Go et al.) with pre-labeled tweets to train the Naive Bayes model, and collected and labeled tweets related to companies in interest to form a smaller set for testing purpose. We tried two approaches for the sentiment analysis task: one depends on lexicons of sentiment-related words, and the other one based on Naive Bayes. We utilized the NLTK Vader Sentiment Analysis tool (Ribeiro et al., 2016) with its pre-
built lexicon for the first approach, and we trained our own model with the scikit-learn tool (Pedregosa et al., 2011) for the second one. For the stock prediction, we implemented a LSTM model (Brownlee, 2017)(Singh, 2018) with only historical prices as predictor as baseline, and implemented four more new models with predictors combining tweet count, daily sentiment score, and weekly sentiment score.

2.1. Data

There were two sources of data (tweets) used in sentiment analysis. We used Sentiment140, which is a large dataset with 1,600,000 labeled tweets in the training set and 498 labeled tweets in the test set. These tweets and labels were used for Naive Bayes model training and testing. The tweets included are of general topics, and the sentiment is scaled in 0, 2, and 4 for negative, neutral, and positive respectively. Since the Naive Bayes model is binomial, neutral tweets from the data set were excluded from training and test. To make sure that the model trained on a general tweet set could also be applied to tweets containing company keywords with a different vocabulary set, we also created a customized small dataset with 454 tweets for test. These tweets containing specified company names, and were labelled manually as negative, neutral, or positive. For the Naive Bayes model only the negative and positive tweets were tested, but all tweets were used for VADER testing, a model does not need training.

We used company stock price as the metric of the business performance since it is easy to obtain, updated daily and also often used as a key indicator of a company valuation and performance. To obtain stock price, we used the End of Day US Stock Prices database provided by Quotemedia in Quandl API (Quandl). We chose adjusting closing price as the metric because it is the price that reflects the true value of a company and accounts for all corporate actions such as stock splits, dividends distributions and rights offerings.

2.2. Collect company related tweets

To obtain the customized test set of tweets, we collected 70 tweets for each of the companies in the list: Tesla, usps, Starbucks, PandaExpress, Chipotle, Uber, Lyft. These companies are chosen because have a focus in customer feedback and it was less likely to have irrelevant tweets.

For generating scores for the second stage of prediction, we further used advanced search of Twitter to define the search time frame and collected tweets containing Tesla, Starbucks, and Chipotle respectively from January 1st 2018 to November 30th 2018. These were used in training and testing of the prediction model.

2.3. Sentiment Analysis

2.3.1. Lexicons of sentiment-related words

The VADER sentiment analysis tool is a pre-made model that is based on lexicons of sentiment-related words. In this approach, each word in the vocabulary set is manually assigned a score for the sensitivity that they usually express. VADER treats a sentence by looking for each of the words included in the sentence in its vocabulary dataset, and assigning corresponding sensitivity score to it. It then generates a compound score for the sentence in the range [-1,1] based on the identified words. With its vocabulary set, this model does not need training, and we tested the customized set on it to see how it performs for these company related tweets.

2.3.2. Naive Bayes

Trying to get a better result than the algorithm based on lexicon, we trained a binomial model based on Naive Bayes. We applied the scikit-learn tool, and the model is trained on the training set of Sentiment140 training set, excluding tweets scaled as 2 (neutral). We then tested the model by comparing the prediction with the labels of the test set. The trained model was also tested on our customized set for two purposes: to guarantee that this model could be used in our specific case about company tweets, and to make the result comparable with the one generated by VADER model.

2.4. Stock Prediction

For stock prediction, we implemented the LSTM model with previous 5 days stock prices as the base case to be compared with. The model used data of 175 trade days from January 1, 2018 to September 18, 2018 for training, and used data of 54 trade days from September 18, 2018 to November 30, 2018 for test. We wanted to explore if the sentiments of company related tweets and the number of tweets collected could be further divided into two prediction models: daily and weekly. Daily score is the average
score of all tweets of that specific day, and weekly score is a moving average score of all tweets of the previous five days. Four new versions of prediction model with various predictors were implemented: model with only daily score, model with only weekly score, model with tweet count and daily score, and model with tweet count and weekly score, and the performances were compared. There were two metrics to measure and compare the performances. The first is the MSE of prediction price, which calculates the error based on difference between predicted prices and actual prices. The second is the trend accuracy, which tested whether the model predicted the trend of price, i.e. go up or down, correctly for the next day.

3. Results

3.1. Sentiment Analysis

We first tested with the VADER sentiment analysis tool, and it labeled 219 wrong out of the 454 manually labeled tweets, resulting in an accuracy rate of 0.5176. This accuracy rate is not optimal so we used it as the baseline model and sought other algorithm that can provide a better performance. For the Naive Bayes Model, since it was trained with the pre-labeled data set, it labeled 85 wrong out of the 359 non-neutral tweets, having an accuracy rate of 0.7632. We also tested the model with our customized test set to further verify that it is applicable. Possibly due to the difference in vocabulary, the accuracy rate has been lowered slightly. Out of the 271 non-neutral tweets, it labeled 81 wrong, resulting in an accuracy rate of 0.7011.

3.2. Stock Prediction

Evaluated from normalized root MSE, the baseline of model without sentiment score already gives good performance with MSE less than 4.5% for each of the companies. By comparing performances, we do not see a significant difference for three of the new models, and only the one with tweet count and daily score as predictors gave a much worse output.

However, evaluated from trend accuracy, the baseline model did not perform well, with accuracy of 43.14%, 50.98%, and 49.02% respectively for Tesla, Chipotle, and Starbucks, worse than pure guessing. Most of the new models again do not mark a significant difference from the baseline, and the one with tweet count and daily score performs even worse than the baseline.

4. Conclusion

By comparing the performance of different models on our customized test set, we see a significant advantage for the Naive Bayes Model. We believe the disadvantage for VADER is that by tokenizing the sentence, it considered every word separately without taking into account the context. For the second stage of the project, by comparing the result of prediction for Tesla, Starbucks, and Chipotle, we noticed that including the sentiment score as a predictor for stock price prediction has no significant impact on the prediction results of both MSE and trend accuracy. Therefore, we conclude that feedback in social media have no significant influence on company stock prices. However, this negative result might be due to the small set size of company related tweets that we collected. We also didn’t consider the difference in the influence that each account has and just assigned each tweet the same weight. This can also limit how the sentiment score can actually reflect customer feedback.

5. Acknowledgement

Thanks to our instructor Mahmoud Mostapha for providing insight and suggestions regarding our project.

References

Proc. The Effect of Sentiment on Stock Price Prediction.(Vanstone B.J., Gepp A., Harris G. (2018))In In-
Figure 3. Prediction of model with tweet count and weekly score

---


Go, Alec, Bhayani, Richa, and Huang, Lei. Twitter sentiment classification using distant supervision.

Kouloumpis, Efthymios, Wilson, Theresa, and Moore, Johanna D. Twitter sentiment analysis: The good the bad and the omg! 2011.


