Abstract

Relationship Between Financial Tweets and Stock Market Data

Panni Homolya

2023

Nowadays, Twitter has a wide range of posts, on various topics, including the latest issues and events in the economy and business. It is often stated that posts may impact the stock market, or that stock market movements will start to generate positive or negative tweets. In the thesis, we investigate the relationship between tweet sentiments and the stock market, and create a *Random Forest* model to predict the logreturns calculated from the close values of the stock market.

At the beginning of the thesis, we review related studies in this area. After, we present the Twitter and stock market datasets used in the thesis, which include data from the following large impact companies: *Tesla*, *Apple*, *Amazon*, *GOOG*, *GOOGL* and *Microsoft*. We describe mathematical tools used in our work, such as *FinBert* language model, *Spearman's rank correlation*, *Granger causality* or *Random Forest* method. Furthermore, we present the *MAE*, *MSE*, *RMSE* and R^2 measures used in the evaluation for validation purposes.

In correlation analysis, we created a dataset that expresses logreturn predictions from tweet sentiments and vice versa. On this dataset we found correlation from daily logreturn value to tweets prediction. Absolute logreturn correlates with tweet volume. We added new features to the data to investigate the long term prediction. Using Granger causality, sentiments Granger causes logreturn in the long run. The logreturn Granger causes sentiments in the short run. Taking the above mentioned results into account, we trained a Random Forest model separately for each company and also created a general model. They showed different predictive performance across companies, with R^2 score tanging from 0.0387 to 0.2796 and 0.2048 for the general model.

The *Random Forest* provides satisfiying results to predict stock market flow for *Tesla*. In the meanwhile it cannot predict well the logreturn of *GOOG*, in that case general model can be used. The general model may not accurately predict all companies, but it can predict logreturns of small impact companies as well, given sufficient data.