Paper 6 | Sentiment Analysis of Twitter Data for Predicting Stock Market Movements by Venkata Sasank Pagolu et. al. – published in 28th October 2016

Abstract

1. Now-a-days social media is perfectly representing the public sentiment and opinion about current events
2. The paper aims to observe how well the changes in stock prices of a company, the rises and falls, are correlated with the public opinions being expressed in tweets about that company
3. The paper has employed two different textual representations: **Word2Vec** and **N-grams**
4. **The overall objective of the paper is to analyse the correlation between stock market movements of a company and sentiments in tweets**. A simple intuition is that positive news and tweets in social media would encourage people to invest in the stocks of that company, resulting in an increase in the stock price of that company
5. **The result shows that there is a strong correlation between the rise and falls in stock prices and the public sentiments in tweets**
    - Specifically, there is a strong correlation between twitter sentiments and the next day stock prices!
6. The keywords used for extracting tweets are very important to accurately capture the public opinions/emotions of a particular stock

Related work

1. Most well-known publication is Bollen
    - Investigated whether the collective mood states of public (derived from twitter feeds) are correlated to the value of Dow Jones Index. They used **Fuzzy NN** for their prediction
    - Their results show that there is a strong correlation between public mood states in twitter and the Dow Jones Index
2. Chen and Lazer
    - Derived investment strategies by observing and classifying the twitter feeds
3. Zhang
    - Found a high negative correlation between mood states and Dow Jones Average Index
4. Brian et. al
    - Investigated the correlation of sentiments of public with stock increase and decreases using Pearson correlation coefficient for stocks

Sentiment Analysis

1. Tweets are classified as positive, negative, and neutral
2. Feature extraction
    - N-gram representation
        - Break down tweets into n-grams, where n denotes the length of the word sequence
        - For example, “Microsoft is launching a new product”

        - A 3-gram word features would be “Microsoft is launching”, “is launching a”, “launching a new”, and “a new product”
        - N-grams for all the tweets form the corpus
        - Therefore, tweet is split into N-grams and the features to the model are a string of 1s and 0s where 1 represents the presence of that N-gram of the tweet in the corpus
    - Word2vec representation
        - Mapping words to 300-dimensional vector representations
        - Vectors of words can be summed up to yield a resultant vector for any given collection of words
        - Relationships between the words is retained in this vector form
        - For example, the word vectors difference between Rome and Italy is very close to the different between vectors of France and Paris
3. Model training
    - Used random forest algorithm
    - The model trained with word2vec is chosen because of its sustainability of meaning and promising performance over large datasets

Correlation analysis of price and sentiment

- If the previous day stock price is more than the current day stock price, current day is marked with 0, else marked with 1
- This makes the correlation analysis a classification problem
- The total positive, negative, and neutral sentiments in a 3-day period are calculated successively which are used as features for the classifier model and output the next day value of stock: 0 or 1

Results

1. Sentiment analyser results

TABLE II: Sentiment Analysis Results

Machine Learning Algorithm	Word2vec				N-gram			
	Accuracy	Precision	Recall	F-Measure	Accuracy	Precision	Recall	F-Measure
Random Forest	70.18%	0.711	0.702	0.690	70.49%	0.719	0.705	0.694
Logistic Regression	62.42%	0.621	0.624	0.621	57.14%	0.580	0.571	0.574
SMO	62.42%	0.617	0.624	0.618	65.84%	0.658	0.658	0.657

2. Stock price and sentiment correlation results
    - The classifier results show an accuracy value of 69.01% when trained using Logistic regression
    - LibSVM model gave a result of 71.82%
    - These results show a good correlation between stock market movements and the sentiments of public expressed in twitter

Conclusion

1. There is a strong correlation exists between rise/fall in stock prices of a company to the public opinions or emotions about that company on twitter

Future work

1. Explore other social media platforms. For example, **Stocktwits** is a financial communication platform designed solely for sharing ideas and insights of investors