Sentiment analysis on Twitter feeds using successive deviation technique for prediction of stock market shift

Tejas Sathe, Siddhartha Gupta, Shreya Nair, Sukhada Bhingarkar

Abstract– The stock market is fluctuating constantly. The rise and fall in stock prices are seemingly random. However, this is not so. Even a minute happening in the company can have a huge effect on the stock price. As each investor buys and sells the stock, the price rises and falls depending on the sale and purchase, the demand and supply. Whether or not an investor buys a particular company's stock is based on his knowledge and impression of the company. The latter is what we will employ to decide whether or not to buy a certain company's stock at the current price.

There are 6 accepted discrete moods. Millions of people tweet every second. A fairly accurate prediction and analysis of the tweet's underlying mood can be made using sentiment analysis. Each word has a certain grammatical signature that tells us which mood it belongs to. Depending on what the users are feeling about a company as they tweet about it this engine will decide whether or not one should buy stocks of that company.

Keywords– Twitter, Sentiment Analysis, successive deviations, stock market, mood

1. Introduction

Sentiment analysis aims to determine the attitude of a speaker or a writer (in this case the person who is tweeting) with respect to some topic (in this case the stock market). Sentiment Analysis refers to the use of natural language processing,text analysis and computational linguistics to identify and extract subjective information in source materials.[1]

Stock market being a volatile market, it isn't possible to predict the stock market accurately with the help of standard and existing algorithms. As the mood of the market is set by people's emotions towards the market, analyzing the tweets would give us an overview of the people's emotions which would help us predict if the market is in bullish or bearish mood.

A bull market is when the market is showing confidence with the prices going up. A bear market is the opposite of bull market. A bear market shows low confidence and in this market the prices drop.[2]

Stock market plays a crucial role in the economy of a country. They are an integral part of all major economies. They provide unique services and benefits to corporations, individual investors and governments.[3] Billions of dollars are traded on different stock exchanges everyday.[4] Being able to get a foresight of the direction the stock market is moving will be financially beneficial. We will do this for a large number of people. Depending on the overall nature of sentiment (positive or negative) we will attempt to predict whether a particular stock will rise or fall. We will test this against a mock portfolio to establish the validity of the hypothesis.

The remaining paper is organized as follows:

- Section 2 describes our research into all the related work that has already been done in this field.
- Section 3 gives a short explanation of Sentiment Analysis and its relevance in this paper.
- Section 4 describes the algorithm that we propose.
- Section 5 lists several applications of this concept.
- Section 6 describes the simulation that we propose to run to test the engine.
- Section 7 concludes the paper and lists the work that can be done in this field in the future.

2. Related Work

Sentiment Analysis as an area is on the rise. It comes under the Natural Language Processing field. It ranges from document level classification to learning the polarity (positive or negative connotation) of words and phrases.

Turney and Pang performed some of the earliest work in this field. They applied different methods for detecting the polarity of product reviews and movie reviews respectively. This was at the document level. A document's polarity can also be classified, which was done by Pang and Snyder. Pang expanded the basic task of classifying a movie review as either positive or negative to predicting star ratings on either a 3 or a 4 star scale, while Snyder performed an in-depth analysis of restaurant reviews, predicting ratings for various aspects of the given restaurant, such as the food and atmosphere (on a five-star scale). A different method for determining sentiment is the use of a scaling system whereby words commonly

associated with having a negative, neutral or positive sentiment with them are given an associated number on a -10 to +10 scale (most negative up to most positive) and Another research direction is the analysis of *subjectivity* and *objectivity*. This involves the classification of a piece of text as either subjective or objective. This is generally more difficult than polarity determination. The subjectivity of words and phrases may depend on their context and an objective document may contain subjective sentences (e.g., a news article quoting people's opinions).

Yet another analysis model is the *feature/aspect-based sentiment analysis* model. It refers to determining the Modern sentiment analysis can be used for a variety of things. Open source software tools deploy machine learning, statistics, and natural language processing techniques to automate sentiment analysis on large collections of texts, including web pages, online news, internet discussion groups, online reviews, web blogs, and social media. As automated systems cannot account for historical tendencies, a human component to sentiment analysis is essential. Automation can only accurately predict 23% of comments correctly classified by humans. Sometimes, the structure of comments and sentiments is fairly complex. Also, the problem of sentiment analysis is non-monotonic in respect to sentence extension and stop-word substitution (compare *THEY would not let my dog stay in this hotel* vs *I would not let my dog stay in this hotel*). To address this issue a number of rule-based and reasoning-based approaches have been applied to sentiment analysis, including Defeasible Logic Programming.

Thus, all in all, modern sentiment analysis can predict the mood of a statement fairly accurately. There are 6 distinct, discernible moods that a statement can be classified into.

As Twitter imposes a restriction on the number of characters per tweet, this process becomes somewhat simpler as the size of the data is reduced. Various ways of analyzing twitter sentiments have cropped up. Several researchers and companies rely on the analysis of emoticons for prediction mood of the sentence[9].

This mood or connotation has been used in various ways to predict future trends. Godbole et al [8] propose several parameters (subjective as well as objective) to achieve this end.

They define the following terms:

Polarity: Tells us whether the mood/connotation of the sentence is positive or negative.

Subjectivity: Tells us exactly how much sentiment is contained in a particular sentence.

world_polarity = (positive sentiment references) / (total sentiment references)

when a piece of unstructured text is analyzed using Natural Language Processing. Depending on the score, the polarity of that word is determined.

opinions or sentiments expressed on different features or aspects of entities, e.g., of a cell phone, a digital camera, or a bank. A feature or aspect is an attribute or component of an entity, e.g., the screen of a cell phone, or the picture quality of a camera. This problem involves several sub-problems, e.g., identifying relevant entities, extracting their features/aspects, and determining whether an opinion expressed on each feature/aspect is positive, negative or neutral.

entity_polarity, however, is calculated only for a particular day (day_i).

$entity\_polarity_i = (positive\ sentiment\ references_i) / (total\ sentiment\ references_i)$

These two parameters will accurately tell us the overall mood of a time period. [8]

Furthermore, moods can also be guessed using emoticons, which are cartoon faces with human expressions. They are classified into happy and sad emoticons. These are hard coded, and the underlying symbols have no significance in determining the mood depicted by the emoticon.

Happy emoticons: ":-)", ":)", "=)", ":D" etc.
Sad emoticons: ":-(", ":(", "=(", ";(" etc.

These will tell a classifier to understand a sentence's mood as positive or negative based on the accompanying emoticon [7]. This is extremely accurate, as no sane person would post a happy emoticon with a sad sentence or vice versa, or post both types of emoticons in the same sentence.

Twitter is one of the biggest social networks that exists today. People like to post minutiae of their lives on such networks. While some people might argue that this is irritating and a waste of bandwidth, time and who knows what else, many entities certainly use it to their advantage. For example, such details of people's lives gives a rare insight into what they like and dislike, what they subscribe to, where they shop and shop and so much more. This data is analyzed and used by companies to advertize their product to only a select number of people. This gives them a target audience, increasing the odds of their product being sold.This can be easily achieved by Twitter's public domain hashtag data set. People post relevant hashtags at the end of each tweet. These can be used to gauge their interests. This dataset is a subset of the Edinburg Corpus dataset[9].

This can also be used to predict stock market shift, feeding

upon people's satisfaction or discontentedness with a particular company. Chen et al [5] propose using this to determine not only the direction but also the intensity of the shift. Based on this they come up with a suitable investment strategy that they suggest to the user, one in which they feel the user will have maximum monetary gain.

3. Sentiment Analysis

The entire tweet will be parsed to determine the overall mood of the user as he tweets. For this we will be using a text file generated by SentiWordNet, which gives positive and negative scores for a group of words. These groups, called 'synsets' (synonym sets), depict the same overall mood. The Objectivity Score (OS) of each word can be calculated as 1 – (positive score + negative score). Thus, the tweet will be stripped of punctuations and a list words will be generated. Again, applying a combination of NLP and Regex the SMS slang will be filtered to give proper words. Then, each word's Objectivity Score will be calculated using the aforementioned formula. A summation of these OSs will reveal whether the overall mood is positive or negative. These characterizations can also be called as 'happy' or 'sad'. This is for just one tweet. Hundreds of such tweets will be gathered at an instant and the overall mood of the group of tweets will be analyzed. Depending on whether it is 'happy' or 'sad', the user of this engine will be advised to or against buying the stock of one company.

4. Interpretation determination by successive deviation calculation

A word can have several interpretations. Each one, in turn, has its own Objectivity Score. Thus, the most relevant interpretation for that sentence has to be chosen. Only then can the mood of the sentence be accurately determined. If a different interpretation is considered it may change the overall mood of the sentence.

4.1 The Proposed Algorithm

Thus, the following algorithm outlines the selection of the correct interpretation:

1. Separate the tweet into a list of words.

2. Get the Objectivity Score for each of the word's interpretations.

3. Compare each pair of the Objectivity Scores of the first two words and prepare a list of deviations.

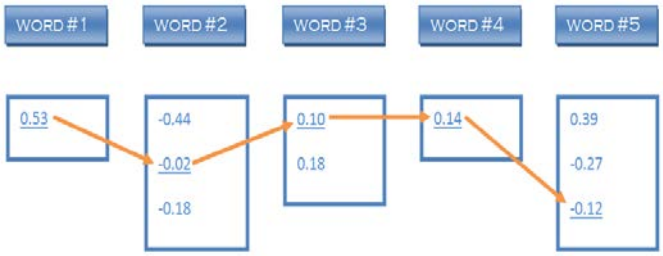
4. Choose the pair with the minimum deviation.

5. Consider the interpretation of the second word from the pair to calculate the deviation with the third word.

6. Repeat this process till the word list is exhausted.

7. Add all the Objectivity Scores to get the overall mood.

8. Perform this for all the tweets and sum the OSs up to get the overall mood of the group of tweets and predict the movement of the market.



As can be seen, each word has one or more interpretations, each of which has its own objectivity score. The deviation between 0.53 and -0.02 is the minimum. Thus, -0.02 is chosen as the correct one. Again, with respect to -0.02, 0.10 has least deviation. Hence, that one is chosen as the correct interpretation. Continuing this process for each word, we get the concerned objectivity score.

5. Applications

5.1 Investment Strategies

Several investment strategies can be suggested based on the results of the above algorithm.

Simple: This strategy will simply advise the user to buy as many stocks if the predicted movement of the market is positive. Otherwise, it would tell the user to just hold on to the money and check the following day to see if he can invest then.

Complex: Here, we would also try to predict by how much the market would shift, instead of just the direction it would shift in. This would be done based on the number of 'happy' v/s the number of 'sad' tweets. We can then come up with scheme on exactly how much to invest depending on the amount of shift the market will encounter.

5.2 Some Benchmark Investment Strategies

Default: In this strategy the investor will simply keep

buying shares as long his money doesn't run out.

Maximal: This strategy assumes perfect knowledge about future stock prices, general market trends and goings-on of that company. We invest everything we have when we predict a rise in the market and nothing when we think it will fall. This strategy is impossible to implement in real life, as no one can really know whether the market will rise or fall.

6. Simulation

We propose starting with enough money to buy, say 100 shares of a particular company. Depending on the algorithms stated, we would try to predict the direction of shift of the market, and 'invest' in that company (we wouldn't actually invest, just deduct that money from our balance and keep track of money coming in and going out). The only drawback of this approach is that if we had actually bought the shares in the real market, the prices would have been slightly different due to our purchase. But that fluctuation is so small that it can be ignored.

7. Conclusion and future work

Thus, we have surveyed related work on sentiment analysis in general and on Twitter feeds, and have proposed a new algorithm to select the best and the most relevant Objectivity Score among a variety of the same. In the future, we hope to implement an engine that accurately predicts stock market shift based on sentiment analysis. We believe that the accuracy of this prediction will be increased due to the proposed algorithm.

References

[1]Sentiment Analysis: http://en.wikipedia.org/wiki/Sentiment_analysis

[2] What is a Bull and a Bear market?http://content.moneyinstructor.com/693/what-bull-bear-market.html

[3]The Role of the Stock Market: http://www.updown.com/education/article/The-Role-of-the-Stock-Market

[4]Daily Market Summary: http://www.nasdaqtrader.com/Trader.aspxid=DailyMarketSummary

[5] Ray Chen, Marius Lazer, "Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement"

[6] SentiWordNet. An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. http://sentiwordnet.isti.cnr.it/

[7] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10, Valletta, Malta, European Language Resources Association ELRA,(May 2010*)

[8] Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena, "Large-Scale Sentiment Analysis for News and Blogs", *Proceedings of the International Conference on Weblogs and Social Media ICWSM,* (*2007*)

[9] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", *Proceedings of the International Conference on Weblogs and Social Media ICWSM (2011)*