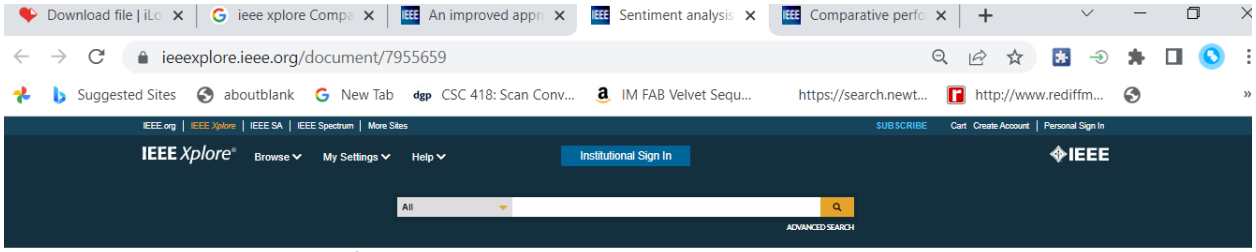


Sentiment analysis of Twitter data for predicting stock market movements

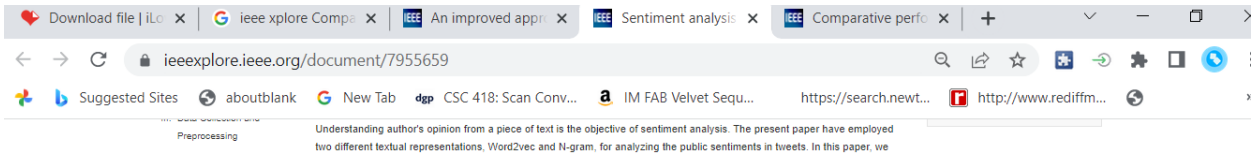
Publisher: IEEE

Cite This | PDF

Venkata Sasank Pagolu ; Kamal Nayan Reddy ; Ganapati Panda ; Babita Majhi All Authors

Abstract:
Predicting stock market movements is a well-known problem of interest. Now-a-days social media is perfectly representing the public sentiment and opinion about current events. Especially, Twitter has attracted a lot of attention from researchers for studying the public sentiments. Stock market prediction on the basis of public sentiments expressed on Twitter has been an intriguing field of research. Previous studies have concluded that the aggregate public mood collected from Twitter may well be correlated with Dow Jones Industrial Average Index (DJIA). The thesis of this work is 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. Understanding author's opinion from a piece of text is the objective of sentiment analysis. The present paper have employed two different textual representations, Word2vec and N-gram, for analyzing the public sentiments in tweets. In this paper, we have applied sentiment analysis and supervised machine learning principles to the tweets extracted from Twitter and analyze the correlation between stock market movements of a company and sentiments in tweets. In an elaborate way, positive news and tweets in social media about a company would definitely encourage people to invest in the stocks of that company and as

More Like This
Stock Price Prediction using Machine Learning and Sentiment Analysis
2021 2nd International Conference for Emerging Technology (INCET)
Published: 2021
A Novel Stock Price Prediction Scheme from Twitter Data by using Weighted Sentiment Analysis
2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)
Published: 2022
Show More



Understanding author's opinion from a piece of text is the objective of sentiment analysis. The present paper have employed two different textual representations, Word2vec and N-gram, for analyzing the public sentiments in tweets. In this paper, we have applied sentiment analysis and supervised machine learning principles to the tweets extracted from Twitter and analyze the correlation between stock market movements of a company and sentiments in tweets. In an elaborate way, positive news and tweets in social media about a company would definitely encourage people to invest in the stocks of that company and as a result the stock price of that company would increase. At the end of the paper, it is shown that a strong correlation exists between the rise and falls in stock prices with the public sentiments in tweets.

Published in: 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)

Date of Conference: 03-05 October 2016
INSPEC Accession Number: 16980533
Date Added to IEEE *Xplore*: 26 June 2017
DOI: 10.1109/SCOPES.2016.7955659
ISBN Information:
Publisher: IEEE
Conference Location: Paralakhemundi, India

I. Introduction

Earlier studies on stock market prediction are based on the historical stock prices. Later studies have debunked the approach of predicting stock market movements using historical [illegible] largely fluctuating. The efficient market hypothesis (EMH) states that financial market mo[illegible]events and product releases and all these factors will have a significant impact on a company's stock value [2], because of the [illegible]ying unpredictability in news and current events, stock market prices follow a random walk pattern and cannot be predicted with more than 50% accuracy [1].

Sentiment Analysis of Twitter Data for Predicting Stock Market Movements

Venkata Sasank Pagolu
School of Electrical Sciences
Computer Science and Engineering
Indian Institute of Technology,
Bhubaneswar, India 751013
Email: vp12@iitbbs.ac.in

Kamal Nayan Reddy Challa
School of Electrical Sciences
Computer Science and Engineering
Indian Institute of Technology,
Bhubaneswar, India 751013
Email: kc11@iitbbs.ac.in

Ganapati Panda
School of Electrical Sciences
Indian Institute of Technology
Bhubaneswar, India 751013
Email: gpanda@iitbbs.ac.in

Babita Majhi
Department of Computer Science and IT
G.G Vishwavidyalaya, Central University
Bilaspur, India 495009
Email: babita.majhi@gmail.com

***Abstract*—Predicting stock market movements is a well-known problem of interest. Now-a-days social media is perfectly representing the public sentiment and opinion about current events. Especially, twitter has attracted a lot of attention from researchers for studying the public sentiments. Stock market prediction on the basis of public sentiments expressed on twitter has been an intriguing field of research. Previous studies have concluded that the aggregate public mood collected from twitter may well be correlated with Dow Jones Industrial Average Index (DJIA). The thesis of this work is 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. Understanding author's opinion from a piece of text is the objective of sentiment analysis. The present paper have employed two different textual representations, Word2vec and N-gram, for analyzing the public sentiments in tweets. In this paper, we have applied sentiment analysis and supervised machine learning principles to the tweets extracted from twitter and analyze the correlation between stock market movements of a company and sentiments in tweets. In an elaborate way, positive news and tweets in social media about a company would definitely encourage people to invest in the stocks of that company and as a result the stock price of that company would increase. At the end of the paper, it is shown that a strong correlation exists between the rise and falls in stock prices with the public sentiments in tweets.**

Keywords: Sentiment Analysis, Natural Language Processing, Stock market prediction, Machine Learning, Word2vec, N-gram

I. Introduction

Earlier studies on stock market prediction are based on the historical stock prices. Later studies have debunked the approach of predicting stock market movements using historical prices. Stock market prices are largely fluctuating. The efficient market hypothesis (EMH) states that financial market movements depend on news, current events and product releases and all these factors will have a significant impact on a company's stock value [2]. Because of the lying unpredictability in news and current events, stock market prices follow a random walk pattern and cannot be predicted with more than 50% accuracy [1].

With the advent of social media, the information about public feelings has become abundant. Social media is transforming like a perfect platform to share public emotions about any topic and has a significant impact on overall public opinion. Twitter, a social media platform, has received a lot of attention from researchers in the recent times. Twitter is a micro-blogging application that allows users to follow and comment other users thoughts or share their opinions in real time [3]. More than million users post over 140 million tweets every day. This situation makes Twitter like a corpus with valuable data for researchers [4].Each tweet is of 140 characters long and speaks public opinion on a topic concisely. The information exploited from tweets are very useful for making predictions [5].

In this paper, we contribute to the field of sentiment analysis of twitter data. Sentiment classification is the task of judging opinion in a piece of text as positive, negative or neutral.

There are many studies involving twitter as a major source for public-opinion analysis. Asur and Huberman [6] have predicted box office collections for a movie prior to its release based on public sentiment related to movies, as expressed on Twitter. Google flu trends are being widely studied along with twitter for early prediction of disease outbreaks. Eiji et al. [11] have studied the twitter data for catching the flu outbreaks. Ruiz et al. [7] have used time-constrained graphs to study the problem of correlating the Twitter micro-blogging activity with changes in stock prices and trading volumes. Bordino et al. [8] have shown that trading volumes of stocks traded in NASDAQ-100 are correlated with their query volumes (i.e., the number of users requests submitted to search engines on the Internet). Gilbert and Karahalios [9] have found out that increases in expressions of anxiety, worry and fear in weblogs predict downward pressure on the S&P 500 index. Bollen [10] showed that public mood analyzed through twitter feeds is well correlated with Dow Jones Industrial Average (DJIA).

All these studies showcased twitter as a valuable source and a powerful tool for conducting studies and making predictions.

Rest of the paper is organized as follows. Section 2 describes the related works and Section 3 discusses the data portion demonstrating the data collection and pre-processing part. In Section 4 we discuss the sentiment analysis part in our work followed by Section 5 which examines the correlation part of extracted sentiment with stocks. In Section 6 we present the results, accuracy and precision of our sentiment analyzer followed by the accuracy of correlation analyzer. In Section 7 we present our conclusions and Section 8 deals with our future work plan.

II. Related Work

The most well-known publication in this area is by Bollen [10]. They investigated whether the collective mood states of public (Happy, calm, Anxiety) derived from twitter feeds are correlated to the value of the Dow Jones Industrial Index. They used a Fuzzy neural network for their prediction. Their results show that public mood states in twitter are strongly correlated with Dow Jones Industrial Index. Chen and Lazer [12] derived investment strategies by observing and classifying the twitter feeds. Bing et al. [15] studied the tweets and concluded the predictability of stock prices based on the type of industry like Finance, IT etc. Zhang [13] found out a high negative correlation between mood states like hope, fear and worry in tweets with the Dow Jones Average Index. Recently, Brian et al. [14] investigated the correlation of sentiments of public with stock increase and decreases using Pearson correlation coefficient for stocks. In this paper, we took a novel approach of predicting rise and fall in stock prices based on the sentiments extracted from twitter to find the correlation. The core contribution of our work is the development of a sentiment analyzer which works better than the one in Brian's work and a novel approach to find the correlation. Sentiment analyzer is used to classify the sentiments in tweets extracted.The human annotated dataset in our work is also exhaustive. We have shown that a strong correlation exists between twitter sentiments and the next day stock prices in the results section. We did so by considering the tweets and stock opening and closing prices of Microsoft over a year.

III. Data Collection and Preprocessing

A. Data Collection

A total of 2,50,000 tweets over a period of August 31st, 2015 to August 25th,2016 on Microsoft are extracted from twitter API [16]. Twitter4J is a java application which helps us to extract tweets from twitter. The tweets were collected using Twitter API and filtered using keywords like $ MSFT, # Microsoft, #Windows etc. Not only the opinion of public about the company's stock but also the opinions about products and services offered by the company would have a significant impact and are worth studying. Based on this principle, the keywords used for filtering are devised with extensive care and tweets are extracted in such a way that they represent the exact emotions of public about Microsoft over a period of time. The news on twitter about Microsoft and tweets regarding the product releases were also included. Stock opening and closing prices of Microsoft from August 31st, 2015 to August 25th, 2016 are obtained from Yahoo! Finance [23].

B. Data Pre-Processing

Stock prices data collected is not complete understandably because of weekends and public holidays when the stock market does not function. The missing data is approximated using a simple technique by Goel [17]. Stock data usually follows a concave function. So, if the stock value on a day is x and the next value present is y with some missing in between. The first missing value is approximated to be (y+x)/2 and the same method is followed to fill all the gaps.

Tweets consists of many acronyms, emoticons and unnecessary data like pictures and URL's. So tweets are preprocessed to represent correct emotions of public. For preprocessing of tweets we employed three stages of filtering: Tokenization, Stopwords removal and regex matching for removing special characters.

1) Tokenization: Tweets are split into individual words based on the space and irrelevant symbols like emoticons are removed. We form a list of individual words for each tweet.

2) Stopword Removal: Words that do not express any emotion are called Stopwords. After splitting a tweet, words like a,is, the, with etc. are removed from the list of words.

3) Regex Matching for special character Removal: Regex matching in Python is performed to match URLs and are replaced by the term URL. Often tweets consists of hashtags(#) and @ addressing other users. They are also replaced suitably. For example, #Microsoft is replaced with Microsoft and @Billgates is replaced with USER. Prolonged word showing intense emotions like cooooooooool! is replaced with cool! After these stages the tweets are ready for sentiment classification.

IV. Sentiment Analysis

Sentiment analysis task is very much field specific. There is lot of research on sentiment analysis of movie reviews and news articles and many sentiment analyzers are available as an open source. The main problem with these analyzers is that they are trained with a different corpus. For instance, Movie corpus and stock corpus are not equivalent. So, we developed our own sentiment analyzer.

Tweets are classified as positive, negative and neutral based on the sentiment present [18]. 3,216 tweets out of the total tweets are examined by humans and annotated as 1 for Positive, 0 for Neutral and 2 for Negative emotions. For classification of nonhuman annotated tweets a machine learning model is trained whose features are extracted from the human annotated tweets.