Stock Market Movements Using Twitter Sentiment Analysis

Nehal Shah

Department of Computer Engineering
Swaminarayan College of Engineering & Technology
Saij-Kalol, Gujarat, India

***Abstract*: in modern era, the utilization of social media has reached unprecedented levels. Among all social media, Twitter is such a well-liked micro-blogging service, that permits users to share short messages in real time concerning events or categorical own opinion. During this paper, we have a tendency to examine the effectiveness of varied machine learning techniques on retrieved tweet corpus. We have a tendency to apply machine learning model to predict tweet sentiment likewise as notice the correlation between twitter sentiment and stock costs. We have a tendency to accomplish this by mining tweets mistreatment Twitter's search API and method it for additional analysis. To work out tweet sentiment, we have a tendency to check the effective 2 machine learning techniques: Naïve mathematician classification and Support vector machines. By evaluating every model, we have a tendency to discovered that support vector machine provides higher accuracy tho' cross validation. When predicting tweet sentiment, we've got mined stock historical information mistreatment Yahoo finance API. We've got designed feature matrix for exchange prediction mistreatment positive, negative, neutral and total sentiment score and stock worth for every day. We've got applied same machine learning rule to work out correlation between tweet sentiments and exchange costs and analyzed however tweet sentiments directly correlates with exchange costs**

***Keywords:* Stock market analysis; sentiment analysis; twitter; microblogging; prediction**

I. INTRODUCTION

In modern era, vast amount of data transmitted online through different social media channels. Data contains information about virtually every topic. Twitter is microblogging platform where 100 million user logins daily and more than 500 million tweets are sent per day. Each tweet has maximum of 280 characters hence more than 70 billion characters generated per day only on twitter. Though each tweet may not valuable, we can extract important data that provide valuable insight about public mood and sentiment score on certain topic.

Recently, in field of computing it's a noteworthy topic to gather large quantity of knowledge from social web site, method it and running knowledge analysis on retrieved knowledge to extract relevant data. Just in case of twitter, we've collected great deal live knowledge, method tweets and generated feature matrix to use machine learning formula. In the stated application, historical finance data is also required to make prediction about stock prices. There are bunch of factors that are involved in prediction of stock prices as well as tweet sentiments. In this paper, we take look at two different machine learning algorithms that mostly use for classification and examine their effectiveness on twitter corpus. We have used implemented techniques in two different ways: 1) To predict tweet sentiment score based on classification co-efficient and 2) The correlation between tweet sentiment and stock market prices.

II. PROBLEM STATEMENT

- How to utilize social media to assess market sentiment and predict the behavior of stock of certain company, market and stock indexes to identify an opportunity for trading?
- In order to design feasible system for identifying an opportunity for trading based on user sentiment. We have used twitter as our social media and analyze it to determine behavior of following stocks: Google, Yahoo, Apple, Microsoft and Goldman Sachs.
- We have chosen twitter as our social media because of its real time facet and importance of real time decision in trading. We have preferred to mine finance historical data from yahoo finance.
- The five stocks were determined based on opinion and mindshare on twitter, ensuring that we will have enough data model how their social data affect their stock.
- Last is to learn and investigate how machine learning techniques can be used to identify trends.

III. ALGORITHM

Classification is an interesting problem in machine learning. In our case we need to construct classifier that is trained our "positive", "negative" or "neutral" labeled tweet corpus.

By using implemented classifier, we have labeled future tweets as either "positive", "negative" or "neutral" based on trained tweets features. In this project, we examine two common classifiers used for text classification: Naïve Bayes Bernoulli and Support Vector Machines models.

In addition, we have implemented Naïve Bayes Classifier. We use it for stock market prediction. We have correlated stock tweets statics and stock prices and classify it for certain day. Train classifier will use to predict stock prices for specific company for next upcoming days.

A. Naïve Bayes Classifier

- A Naïve Bayes classifier is based on Bayes rule and simple form of Bayesian network. It is a simple probabilistic model relies on the assumption of feature independent in order to classify data.
- The algorithm assumes that each feature is independent of presence or absence of other feature in input data as this assumption is known as 'naïve'.
- Bayes classifier use Bayes theorem, which is

$$P(c|F) = \frac{P(F/c)P(c)}{P(F)}$$

P(c/F) = probability of feature F being in class c
P(F/c) = probability of generating feature F given by class c
P(c) = probability of occurrence of class c
P(F) = probability of feature F occurring

- In above context, we are looking for class c, so we find probable class given for given feature, F. Denominator does not depend on class so we treat it as a constant. Numerator depends on class so we focus to determine the value of P(F/c). For a class c_j, the features are conditionally independent of each other, hence

$$P\left(f1, f2 \ldots \frac{fn}{cj}\right) = \prod_i P\left(\frac{fi}{cj}\right)$$

From these we classify tweet statistics with label c* with a maximum posterior decision taking the most probable label from all labels C.

$$c* = arg \max_{cj \in C} P(cj) \prod_i P(fi/cj)$$

The Naïve Bayes is very simple and its conditional independence assumptions are not realistic in real world. However, we have performed it multiple times and it gives better accuracy for stock prediction.

B. Naïve Bayes Bernoulli

- There are two different type of NB classifier named as Multinomial model and Bernoulli model. In this section, we have used Bernoulli model.
- It is equivalent to binary independence model which generates an indicator for each feature of tweet, either 1 indicating presence of feature in tweet or 0 indication the absence of feature in tweet. The Bernoulli model has the same time complexity as the multinomial model.
- In our scenario, we have extract feature set from twitter corpus. The extracted feature will be check in twitter and generates feature data for tweet sentiment analysis.
- If feature exist in tweet, then it is considering as 1 and if feature does not exist in tweet then it is considering as 0. The generated train data is used for classification and trained classify model will be used to predict tweet sentiment.

TRAINBERNOULLINB(C,T)
1. V ←EXTRACTFEATURE(T)
2. N ←COUNT TWEETS(T)
3. **FOR EACH** C ∈ C
4. **DO** N ←COUNTTWEETSINCLASS(T,C)
5. prior[c] ← Nc/N
6. **for each**t∈ V
7. **do** N_{ct}← CountTweetsInClassContainsTerm(T,c, t)
8. condprob[t][c] ← (N_{ct} +1)/ (N_c +2)
9. **return** V, prior, condprob

ApplyBernoulliNB(C,V, prior, condprob)
1. Vd← ExtractTermFromTweet(T)
2. **for each** c ∈ C
3. **do**score[c] ←log prior[c]
4. **for each**t∈ V
5. **do**if t ∈ Vd
6. **then** score[c] += log condprob[t][c]
7. **else**score[c] += log (1-condprob[t][c])
9. **return**arg max $c \in C$score[c]

The Bernoulli model estimates P(t/c) as the fraction of class c that contains term t. When classifying a test tweets, the Bernoulli uses binary occurrence information, ignoring number of occurrences.

C. Multiclass support vector machine

Multiclass SVMs classify an input vector $x \in R^d$ into one of k classes using the following simple rule:

$$\hat{y} = \underset{m \in [k]}{\operatorname{argmax}} \boldsymbol{w}_m^{\mathrm{T}} \boldsymbol{x}.$$

Equation [1]

Each vector $\omega_m \in R^d$ can be thought as prototype representing m^{th}class and the inner product as the score of the m^{th}class with respect to x. Hence above equation chooses the class for highest score. Given n training instances $x \in R^d$ and associates labels $y \in [k]$, the multiclass SVM formulation estimates $\omega_{1,\ldots,} \omega_k$ by solving following problem.

$$\underset{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_k}{\text{minimize}} \ \frac{1}{2}\sum_{m=1}^{k} \|\boldsymbol{w}_m\|^2 + C\sum_{i=1}^{n}\left[1 + \max_{m \neq y_i} \boldsymbol{w}_m^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{w}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i\right]$$

Equation [2]

Where C>0 is regularization parameter and [u]= 0 if u >0 and u otherwise. Intuitively Equation (2) means that, for each training instance, we suffer no loss if the score of the correct class is larger than the score of "closest" class by least 1. In this paper, we assume that $\|x_i\| > 0$ since x_i with $\|x_i\| = 0$. The dual of equation is given by [3], [5]

$$\begin{aligned} \underset{\boldsymbol{\alpha}}{\text{minimize}} \quad & f(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{m=1}^{k} \|\boldsymbol{w}_m(\boldsymbol{\alpha})\|^2 + \sum_{i=1}^{n}\sum_{m=1}^{k} \Delta_i^m \alpha_i^m \\ \text{subject to} \quad & \alpha_i^m \le C_i^m \quad \forall i \in [n] \quad \forall m \in [k] \\ & \sum_{m=1}^{k} \alpha_i^m = 0 \quad \forall i \in [n], \end{aligned}$$

Equation [3]

The primal-dual relationship is given by

$$\boldsymbol{w}_m(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i^m \boldsymbol{x}_i \quad \forall m \in [k].$$

The gradient of f plays an important role and is given by

$$g_i^m = \frac{\partial f}{\partial \alpha_i^m} = \boldsymbol{w}_m(\boldsymbol{\alpha})^{\mathrm{T}}\boldsymbol{x}_i + \Delta_i^m \quad \forall i \in [n] \ \forall m \in [k].$$

Equation [4]

Following [3], we have an optimal solution if and only if $v_i = 0 \ \forall i \in [n]$, where

$$v_i = \max_{m \in [k]} g_i^m - \min_{m \in [k]: \alpha_i^m < C_i^m} g_i^m \quad \forall i \in [n].$$

Equation [5]

The larger v_i, the more $\alpha^1 .. \alpha^k$ violate the optimality conditions. In general, give tolerance parameter ∈, we can stop an optimization algorithm if $v_i < \in, \forall i \in [n]$.

DUAL COMPOSTION FOR MULTICLASS SVMs

The main idea of dual decomposition methods is to update at every iteration a small subset of dual variable, keeping all others fixed. Since the variables $\alpha^1 .. \alpha^k$ associated with x_i are tied by an equality constraint, a natural choice for decomposition methods is to update these variables as a block. Our goal is to find $\delta_I \in R^k$ such that the update $\alpha i \leftarrow \alpha i + \delta_I$ maximizes the decrease of the dual objective. Then restricted problem is

$$\begin{aligned} \underset{\boldsymbol{\delta}_i}{\text{minimize}} \quad & \hat{f}(\boldsymbol{\delta}_i) = \frac{\|\boldsymbol{x}_i\|^2}{2}\|\boldsymbol{\delta}_i\|^2 + \boldsymbol{g}_i^{\mathrm{T}}\boldsymbol{\delta}_i \\ \text{subject to} \quad & \boldsymbol{\delta}_i \le \boldsymbol{C}_i - \boldsymbol{\alpha}_i \\ & \boldsymbol{\delta}_i^{\mathrm{T}}\mathbf{1} = 0, \end{aligned}$$

Algorithm: Dual decomposition for multiclass SVMsInput: {(xi, yi)} n i=1, C > 0, > 0
Initialize α ← 0 and wm ← 0 ∀m ∈ [k]
Shuffle data
repeat
vmax ← −∞
fori∈ [n] **do**

Compute violation vi by Eq. (5)
vmax ← max(vmax, vi)
if vi > 0 **then**
Find δi by solving Eq. (6)
αi ← αi + δi and wm ← wm + δ m i xi ∀m ∈ [k]
end if
end for
until vmax ≤ ∈
Output: α and w1, . . . , wk

IV. Literature Survey

Zhang et al and Gilbert et al [1] assessed the bloggers sentiment from live blog post journal within the polarity of anxiety, concern and worry. They need used Monte Carlo simulation to point out the stock movement in S&P five hundred index. Sprengerset al [2] evaluated stocks from S&P one hundred companies to correlate stock discussions and tweet options that contain Ticker symbols.

Rakhi Batra et al., [3] they implemented the idea by collecting sentiment data and stock price market data and built an SVM models for prediction and in the last they measured the prediction accuracy.

Numerous studies showed that using a mix approach can improve classification scheme [5]. The most common use of sentimental analysis is analyzing of twitter tweets and analyzing the top trends in marketplace [6] Sentimental analysis is also used in sales forecast of a product by examining tweets, weather forecast from tweets and posts from face book[7]. Sentimental analysis had been conducted on stock linked tweets which were collected for a period of 6-month [8]. In order to reduce noise, selection of tweets containing tags of top 100 companies was considered. Each tweet was classified using a Naive Bayes method and a set of 2,500 tweets were trained. Results displayed that sentiment indicators are related with unusual returns and stock volume is linked with trading volume [9]

Jasmina Smailović1, 2 et al., [4] investigates whether Twitter feeds are a suitable data source for predictive sentiment analysis. Financial tweets of eight companies (Apple, Amazon, Baidu, Cisco, Google, Microsoft, Netflix and RIM) were analysed. The study indicates that changes in the values of positive sentiment probability can predict a similar movement in the stock closing price in situations where stock closing prices have many variations or a big fall. Furthermore, the introduced SVM neutral zone, which gave us the ability to classify tweets also into the neutral category, in certain situations proved to be useful for improving the correlation between the opinionated tweets and the stock closing price.

Pang and Lee proposed a novel machine learning method that applies text categorization techniques to just the subjective portions of the document. Extracting these portions can be implemented using efficient techniques for finding minimum cuts in graphs. They used min cut to improve the classification of a sentence into either ‘objective’ or ‘subjective’, with the assumption that sentences close to each other tend to have the same class[10].

V. CONCLUSION

After analyzing tweet sentiment, we will predict stock market prices of particular companies and also try to predict sector wise up and down fall by fetching data from Bombay Stock Exchange site.

VI. ACKNOWLEDGEMENT

The authors wish to thank almighty, parents and well-wishers for their ceaseless bolster, favors and support.

References

[1] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," *Artificial Intelligence*, 2010.

[2] X. Zhang, H. Fuehres and P. A. Gloor, "Predicting stock

[3] **Stock** Prediction Using Twitter Sentiment Analysis
http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf

[4] Jasmina Smailović1,2, Miha Grčar1, Nada Lavrač1, and Martin Žnidaršič1,” Predictive Sentiment Analysis of Tweets: A Stock Market Application” 2013 Springer

[5] M.P.RajaKumar and Dr. V. Santhi. A Comparison of stock Trend Prediction using Accuracy riven neural network variance.Proc. Of INT, conf. on Control, Commmunication and power engineering, 2012, pp. 73-79.

[6] Hui Song,Yingxiang Fan,Xiaoqiang Liu and Dao Tao.

Extracting product features from online reviews for sentimental analysis, Computer Science and Converge Information Technology, 2014, pp. 741-750.

[7] Xinzhi Wang and Xiangfeng Luo.Sentimental Space Based Analysis of User Personalized Sentiments. Ninth International Conference on Semantics, Knowledge and Grids, 2013, pp. 151- 156.

[8] Lu Yonghe and Chen Jianhua. Public Opinion Analysis of Microblog Content, Proc.Of International Con. On Information Science and application, 2014, pp. 1-5.

[9] Sprenger and Webye.Sentiment Analysis of Stock Market News with Semi-supervised Learning, International Conf. On Computer and Information Science, 2012, pp. 325-328.

[10] Pang and L.Lee.2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In ACL-2004.

[11] Large-scale Multiclass Support Vector Machine Training via Euclidean Projection onto the Simplex http://www.mblondel.org/publications/mblondel-icpr2014.pdf

[12] Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation http://apps.cs.utexas.edu/tech_reports/reports/tr/TR-2124.pdf

[13] Market indicators through twitter," 2009 Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis. http://ieeexplore.ieee.org.ezproxy.gl.iit.edu/stamp/stamp.jsp?tp=&arnumber=6753954&tag=1