2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)
Manipal University Jaipur, Sep 28-29, 2019

Time Series with Sentiment Analysis for Stock Price Prediction

Vrishabh Sharma
Department of Information Technology
National Institute of Technology, Karnataka
Surathkal, India
vrishabhsharma22@gmail.com

Rajgauri Khemnar
Department of Information Technology
National Institute of Technology, Karnataka
Surathkal, India
rajgaurikhemnar@gmail.com

Renu Kumari
Department of Information Technology
National Institute of Technology, Karnataka
Surathkal, India
choudharyrenu204@gmail.com

Dr.Biju R Mohan
Department of Information Technology
National Institute of Technology, Karnataka
Surathkal, India
bijurmohan@gmail.com

***Abstract*—Stock price prediction has been a major area of research for many years. Accurate predictions can help investors take correct decisions about the selling/purchase of stocks. This paper aims to predict and gauge stock costs and patterns, utilizing the power of machine learning, content examination and fundamental analysis, to give traders a hands-on tool for keen speculations particularly for the volatile Indian Stock Market. We propose a technique to analyze and predict the stock price with the help of sentiment analysis and decomposable time series model along with multivariate-linear regression.**

***Index Terms*—Stock price prediction, Machine Learning, Time Series, Sentiment analysis, NSE, BSE.**

I. INTRODUCTION

Stocks are good metrics to evaluate the impending and current success of a company. Well informed decisions regarding the purchasing and selling of stocks can help prevent large losses. Numerous approaches have been tried to predict the stock prices. A. Skabar et al.[1] proposed a methodology to determine buy and sell points for financial commodities traded on a stock ex-change using neural networks which can be trained indirectly, using a genetic algorithm based weight optimization procedure. R Nayak[2] researched strategies that outperform the underlying market using a brute force approach to the prediction of stock prices using the predictive power of clustering techniques on stock market data and its ability to provide stock predictions. An approach which performs an evolutionary search of the minimum necessary number of dimensions embedded in the problem using the method of time series prediction was put forth elegantly by T Ferreira et al. in [3].This approach produced promising results by also working on developing a novel clustering strategy. A neural model approach combining financial and economic theory was built in [4] and attempts to identify variables that drive stock prices were tried. In [5], Zitian Wang et al. tried Bayesian graphical method to develop a trend prediction model which considers factors like GDP,PPI and so forth. Heeyoung Lee et al. [6] made further significant contributions to this field by forecasting companies stock price changes in response to financial events reported in 8-K reports.Finally in [7]T. H. Nguyen et al. suggested sentiment analysis on Social Media for Stock Movement Prediction thus taking into account real sentiment which is generated and further guides the market. In this paper, we use a decomposable time series model coupled with sentiment analysis of tweets and multivariate linear regression approach for predicting stock prices. The rest of the paper is organized as follows: Section II deals with the Methodology, Section III deals with the Results, followed by Section IV with Conclusion and Future Work.

II. METHODOLOGY

Figure 1 elucidates the overall methodology proposed in this paper for the prediction of stock prices.We aim that given an indexed company we can with minimum error return a forecasted stock price. This price is a signal as well to whether the market goes up or down. Following are the steps elaborated:

A. Data collection

The data is collected using Twitter API to collect tweets(tweet text, date and time) and using Yahoo Finance data feed to gather stock market data for a particular company. The words appearing in the tweet form the features of the initial dataset.

Since Yahoo API being unreliable due to depreciation,we automatically scrape the same features but from the NSE India(https://www.nseindia.com/) website. For this using BeautifulSoup and NSEPy modules a scraping script is built in python. This provides us accurate real time data within a timeframe specified by input to be trained on.

978-1-7281-1711-9/19/$31.00 ©2019 IEEE

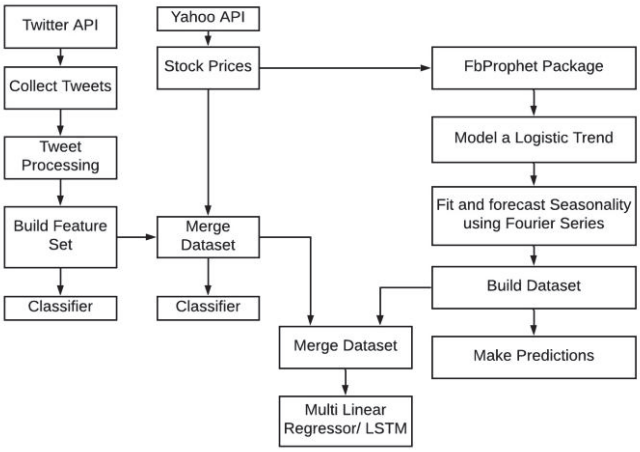

Fig. 1. Flowchart of proposed model

B. Sentiment Analysis

Social media plays important role in predicting the stock market return values. Along these lines, we at that point annexed our information with one more feature - Twitter's Daily Sentiment Score for each organization dependent on the user's tweets about that specific organization and furthermore the tweets on that organization's page .The workflow is as shown in figure 2

To utilize social media to access market sentiment and predict the behavior of stock of certain company, classify polarity of given text at document, sentence or feature level and determine whether opinion in text is positive, negative or neutral and use machine learning algorithm to predict sentiment and find correlation between sentiment and stock prices are the main objectives of this module.

The data is collected using Twitter API to collect tweets(tweet text, date and time) and using Yahoo Finance data feed to gather stock market data for a particular company. The words appearing in the tweet form the features of the initial dataset. The feature set is then build using the words in the vocabulary using the Bag of Words model. Feature selection is performed to select the closest features contributing to the class label. Positive, negative and neutral sentiment score form the class label. This is determined by calculating the total sentiment score of a tweet using the afinn library in python. Tweet Sentiment Analysis is done using Naive Bayes Classification and SVM model.

The data fetched from Yahoo API for each day and the number of positive, negative and neutral tweets posted for the respective company for that day form the features of the next dataset. The class label for each tuple is determined by finding the upward/downward movement of the stock price on each day for the given company. The stock price movement is determined by calculating the difference between closing price and opening price of the stock on each day. Stock market movement is predicted using tweet sentiment analysis and SVM model.

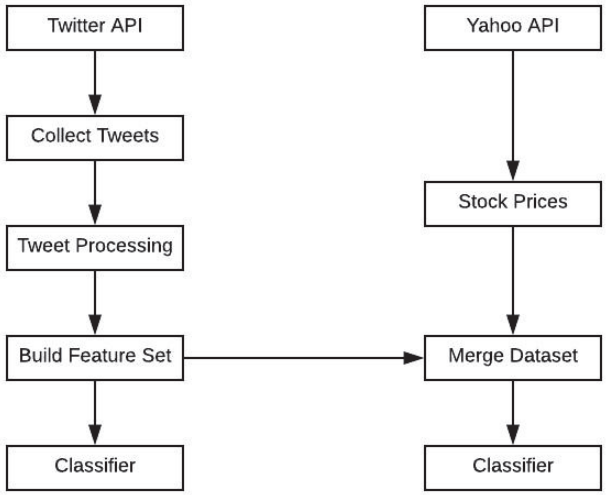

Fig. 2. Sentiment analysis workflow

C. Time Series Decomposition

A decomposable time series model consisting of 3 main model components: trend, seasonality, and holidays is used.

The following equation represents the combined model:

$Y(t) = G(t)+S(t)+H(t)+E(t)$

In our model G(t) represents the trend function, which describes long-term increase or decrease in the data, the nonperiodic changes in the value of the time series. Component S(t) represents periodic changes, which describe how data is affected by seasonal factors such as the weekly, monthly and yearly seasonality. Component H(t) represents the effects of holidays or large events which occur irregularly over one or more days and impact business time series. The error term E(t) represents any aberrant changes which are not included by the model. We make the assumption that E(t) is irreducible and is parametrically normally distributed.

This culminates in the use of a generalized additive model (GAM). GAMs combine the standard generalized linear models with the concept of additive models. Generalized additive models aim at maximizing prediction quality of a dependent variable from many distributions, by estimating nonparametric functions of the predictor variables which are connected to the dependent variable through some link function. The GAM model has the advantage of decomposing and accommodating new components easily, for example identifying a new source of seasonality and including it.

Modelling and fitting are as follows:

Rather than treat is as a standard time series problem, we look towards curve fitting the time series. Thus Trends are modelled by fitting linear and non linear functions of time as components in the GAM. Trends for stock prices represent random exponential trend lines where data values increase or decrease at higher but random rates. For the saturating nonlinear growth a logistic growth model is used. The growth

rate is subject to change as well since growth observed in any period of stock market can fluctuate arbitrarily. Thus points in the series where the growth rate is allowed to change called change points are used. These can be provided based on domain knowledge, scaled for being automatically identified or arbitrarily chosen the number of change points to be included. Seasonality occurs in time series and mimics human behaviour and business decisions. To forecast seasonal effects we must specify seasonality models that are periodic functions of time. Identifying seasonality of a time series can be done through Fourier transforms. A Fourier transform decomposes a signal into all the frequencies that comprise it. We can approximate random smooth seasonal effects with a standard Fourier series with the period of time series as 365.25 days, and include it in the seasonality component of the GAM as a product of a matrix of seasonality vectors and smoothing parameter. This is for each value of time in our data, while the Fourier order is set at 10 to allow modelling high frequency changes. Several instances in real time such as holidays impart changes to a time series which could be predicted. For example, during festivals such as Diwali and Christmas in India public tends to buy a new items at a higher rate, affecting the time series indirectly. Thus we can include a custom list of local and global holidays in the form of a matrix of regressors in our model.

To accomplish this goal we use FBProphet, an open source library published by Facebook. It is based on decomposable (trend+seasonality+holidays) models, provides good accuracy using intuitive parameters and supports inclusion of impact of custom seasonality and holidays. It fits the GAM model by nonlinear optimization algorithm L-BFGS in the probablistic programming language STAN. We use past 10 years data taken from the date of testing on the dataset of open, close, high, low and volume of stocks traded for a particular organization scraped through our script. We forecast a price daily for the timeframe as well as into next 10 days from the date of training. Modelling trends and seasonality we observe a gradual curve-fitting occur with margin for smoothing extra volatility ,should it occur. The model runs for 3180 iterations i.e all days markets are open, as shown in figure 3.

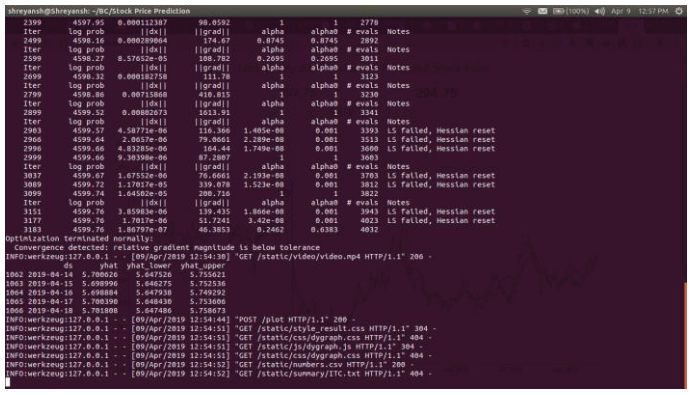

Fig. 3. Time series model run

D. Combining Models

We combined the tweet dataset and time series data to give the final dataset consisting of up/down,no. of positive, negative, neutral tweets ,forecasted prices and actual closing prices. We use a multiple linear regression model to train on one month worth data and predict next 10 day stock price. Working on 50 nse companies we reached a result that has above benchmark accuracy. The timeframe of data used as well as number of days for forecasting can be varied however one month has been specified to keep sentiment more relevant in training the model.Multi Linear regression easily fits non linear data and is thus used in this intuition to combine real time sentiments and forecasted prices provided by in depth training of our time series model.

III. RESULTS

The combined time series-sentiment analysis-MLR model is tested for 50 NSE listed companies most notable Reliance Industries,ITC,Tata Motors and Dr.Reddy's Labs. Price forecasted is checked for error margins against the actual closing prices. Before we elaborate on its accuracy it is important to understand the results of sentiment analysis which are given as accuracy, precision, recall and f-measure metrics for APPLE stock and tweet data in figure 4 and figure 5.

	SENTIMENT PREDICTION FOR AAPL	
	NAIVE BAYES	**MULTICLASS SVM**
ACCURACY	0.7610619469026548	0.7876106194690266
PRECISION	[0.5416666666666666, 0.07692307692307693, 0.9473684210526315]	[0.4583333333333333, 0.6923076923076923, 0.9078947368421053]
RECALL	[0.65, 1.0, 0.782608695652174]	[0.6111111111111112, 0.75, 0.8313253012048193]
F-MEASURE	[0.5909090909090908, 0.14285714285714288, 0.8571428571428571]	[0.5238095238095238, 0.7199999999999999, 0.8679245283018868]

Fig. 4. Results of Sentiment Prediction

STOCK PREDICTION FOR AAPL USING MULTICLASS SVM	
ACCURACY	0.65
PRECISION	[0.7, 0.6]
RECALL	[0.79, 0.75]
F-MEASURE	[0.7422818791, 0.6666666666666666]

Fig. 5. Results of Stock Prediction using Sentiment Analysis

Final forecasted close price results are plotted as a time series over the actual time series in figure 6.

Fig. 6. Final Results

Error measurement statistics play a critical role in tracking forecast accuracy, monitoring for exceptions, and benchmarking forecasting process. The MAPE (Mean Absolute Percent Error) measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error. We use MAPE as our metric between actual closing prices and forecasted prices. Here we compare the results of the proposed method with those of the traditional methods. The results demonstrated in [7]T. H. Nguyen et al. concluded that sentiment analysis on the historical price dataset results in a 60% accuracy to predict stock price movement, whereas, the proposed approach using the aforementioned combined model results in 70% accuracy. Our accuracy as a class-prediction model for UP DOWN or STAY trends is at 70% and above which makes taking decisions in long term trading easier as we can forecast future trend over a large window of time including a span of multiple years. Intraday trading decisions can be made with more confidence as apart from up or down, our accuracy as a continuous value predictor of stock prices using MAPE(mean average percentage error) metric is 1-5%.Thus, our predicted price is always within 1 to 5 percent greater or lesser than of the actual price at that very moment as traded on the market. Thus market volatility is effectively captured and reflected in the forecasts.

IV. Conclusion and Future work

The analysis of our results obtained speaks to us the advanced over the top accuracy in stock price prediction. Our accuracy as a class-prediction model for UP DOWN or STAY trends is at 70% and above which makes taking decisions in long term trading easier as we can forecast future trend over a large window of time including a span of multiple years. Intraday trading decisions can be made with more confidence as apart from up or down, our accuracy as a continuous value predictor of stock prices using MAPE(mean average percentage error) metric is 1-5%.Thus, our predicted price is always within 1 to 5 percent greater or lesser than of the actual price at that very moment as traded on the market. Thus market volatility is effectively captured and reflected in the forecasts. We aim to rectify demerits through a future direction of including annual report analysis published in periodicals by companies of which stock price is to be predicted. This offers a solid fundamental analysis insight into the interested company as we are able to analyse earning, expenditure, management point-of-view, dividends paid out to stakeholders etc. which have a positive correlation in share price changes.

References

[1] A. Skabar, I. Cloete, ”Neural networks, financial trading and the efficient markets hypothesis”, in ACSC 02: Proceedings of the twenty-fifth Australasian conference on computer science, Australian Computer Society, Inc., Darlinghurst, Australia, 2002, pages 241-249

[2] R. Nayak, P.Braak, ”Temporal pattern matching for the prediction of stock prices”, in AIDM ’07:Proceedings of the second international workshop on Integrating artificial intelligence and data mining, Australian Computer Society, Inc, Darlinghurst, Australia, 2007, pages 95103

[3] Tiago A. E. Ferreira, Germano Vasconcelos, Paulo Adeodato, ”A new evolutionary method for time series forecasting”, in ACM proceedings of genetic evolutionary computation conference-GECCO, ACM, Washington, DC, 2005, pages 2221-2222

[4] F. Oliveira, C. Nobre, L. Zarate, ”Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index-Case study of PETR4, Petrobras, Brazil”, Expert Systems with Applications, vol. 40, issue 18, pages 7596-7606, 2013, DOI:10.1016/j.eswa.2013.06.071

[5] Lili Wang, Zitian Wang, ”Stock market trend prediction using dynamical bayesian factor graph”, Expert Systems with Applications, vol. 42, issues 1516, pages 6267-6275, 2015, DOI:10.1016/j.eswa.2015.01.035

[6] Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, Dan Jurafsky, ”On the importance of text analysis for stock price prediction”, in 9th International Conference on Language Resources and Evaluation, LREC 2014 - Reykjavik, Iceland, pages 1170-1175

[7] T. H. Nguyen, K. Shirai, and J. Velcin, ”Sentiment analysis on social media for stock movement prediction”, Expert Systems with Applications, vol. 42, issue 24, pages 9603-9611, 2015, DOI:10.1016/j.eswa.2015.07.052