SENTIMENT ANALYSIS ON RSS NEWS FEED USING NAÏVE BAYES AND SVM

M.SUPRIYA RANI[1], M.V.L.PRATHYUSHA[2], N.SUSMITHA[3], N.HAREESHA[4], P.J.JYOTHI[5]
[1,2,3,4] B.Tech, Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Andhra Pradesh, India
[5]Associate Professor, Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Andhra Pradesh, India

Abstract : As we know that stock market depends on the stock news. This paper is to build a model that predicts news polarity which may affect changes in stock price movement trends .Sentiment analysis is to understand people's actions, feelings, emotions etc. The novelty of our approach is efficient prediction of model that scores emotions from all relevant real time stock news available in public domain. Our paper uses news feeds from Really Simple Syndication (RSS) news feed that has an impact on stock market values. Hence stock market RSS news feed data is collected for a period of time.

We use machine learning Framework for processing data set is available on the RSS news feed. Results of sentiment analysis will be positive or negative or neutral based on the polarity. In our experimental study, Results of sentimental analysis on RSS news feed will be displayed as different sections predicting stock price fluctuations, whether up or down.

IndexTerms_ Sentiment analysis, RSS, Polarity, macine learning.

I. INTRODUCTION

Today, the micro blogging has become a very popular messaging tool between internet users. Millions of users can share their opinions in different aspects of life every day in popular websites like Twitter and Face book. Twitter allows people to create profiles, communicate, and connect with other people on the service. Towards specific product, organization, movies, events, news, issues, services and their attributes, the sentimental analysis is used to obtain the real influence of people. This can be useful in several ways and contains the computer science branches such as Natural Language Processing (NLP), text mining, information theory, machine learning and collecting the training data. The main aim is to identify the sentiment of the tweets or reviews published in the web. First Data streaming is used to merge and access the real-time feeds and archived data for analytics. And then preprocessing of data is done. Later on we perform Chi-Square test on the data to determine the conditional probability of the tweets. Then Sentiment Scoring is done using Naïve Bayes Classifier.

Most people are interested in many websites whose content changes on an unpredictable schedule. Examples of such websites are news sites, community and religious organization information pages, product information pages, medical websites, and weblogs. Repeatedly checking each website to see if there is any new content can be very tedious. Email notification of changes was an early solution to this problem. Unfortunately, when you receive email notifications from multiple websites they are usually disorganized and can get overwhelming, and are often mistaken for spam. RSS is a better way to be notified of new and changed content. Notifications of changes to multiple websites are handled easily, and the results are presented to you well organized and distinct from email. RSS stands for "Really Simple Syndication". It is a way to easily distribute a list of headlines, update notices, and sometimes content to a wide number of people. It is used by computer programs that organize those headlines and notices for easy reading. RSS works by having the website author maintain a list of notifications on their website in a standard way. This list of notifications is called an "RSS Feed". Producing an RSS feed is very simple and hundreds of thousands of websites now provide this feature, including major news organizations like the New York Times, the BBC, and Reuters, as well as many weblogs [1]. For performing Sentimental Analysis on this RSS news feed, we can use several techniques like Naïve Bayes, Support Vector Machine, Decision tree etc.

Related Work

A feature selection for Opinion mining using decision tree is proposed. LVQ type learning models constitute popular learning algorithms due to their simple learning rule, their intuitive formulation of a classifier by means of prototypical locations in the data space, and their efficient applicability to any given number of classes. Movie review features obtained from IMDb was extracted using inverse document frequency and the importance of the word found. Principal component analysis was used for feature selection based on the importance of the work with respect to the entire document. The classification accuracy obtained by LVQ was 75%.However it was observed that the precision for positive opinions was quite low[2].

This research explores the opinion mining on stock market by combining the sensex points of moving average stock level indicator with RSS news feeds which obtained the high accuracy rather than individual sensex calculation [3]. This paper needs to focus on more than one stock level indicator with sentiment analysis.

The impact of sentiments from tweets as well as RSS news feeds is analyzed. It is found that the sentiments from social media along with stock level indicators enhance the quality of prediction. Since our approach is an hybrid approach, the experimental analysis is carried out for hypothesis testing H0 and Ha. Also Ha have shown significant improvement in the precision and correctness when compared with H0. Ha have enough evidence to reject the null hypothesis and accept Alternate hypothesis [4].

This work can further be extended for prediction of buying patterns from customers by incorporating the sentiment mining from various social media such as RSS news feeds, Tweets, LinkedIn and Facebook contents.

sentiment analysis on RSS feeds along with the tweets. We can classify these News and tweets according to area which will help indecision making. It will also help to overcome the weaknesses in particular area. The opinion mining done with RSS feeds and tweets can help a lot to predict the needs of people as well as their views about particular topic

This work can be extended to that due to inclusion of emotions, real time public opinions are not always accurate. So that we have combined the twitter data with RSS feeds to achieve the accuracy [5].

To overcome these, we use Naïve Bayes Algorithm.

PREPROCESSING

Preprocessing is one of the main functionalities.

A. Read RSS News feed

From the relevant web sites, Really Simple Syndication (RSS) news feeds that has an impact on stock market values. Hence stock market RSS news feed data is collected for a period of time. RSS feeder reads the required content such as title, description etc. in the form of XML. These feeds are collected inside this module.

B. Classify RSS News feed

All the collected RSS stock news feeds are stored inside the input sentence module as a whole document.

C. Remove stop word and apply token length files

Spell correction is an important part in sentiment analysis of user-generated content. People use words like 'coooool' and 'hunnnnngry' in order to emphasize the emotion. In order to capture such expressions, we replace the sequence of more than three similar characters by three characters. For example, wooooow is replaced by wooow. We replace by three characters so as to distinguish words like 'cool' and 'cooooool'.

D. Remove weak entities

News consists of various notions of negation. In general, words ending with 'nt' are appended with a not. Before we remove the stopwords 'not' is replaced by the word 'negation' Negation play a very important role in determining the sentiment of the news. This is discussed later in detail.

E. Find term frequency

Numbers

Numbers are of no use when measuring sentiment. Thus, numbers which are obtained as tokenized unit from the tokeniser are removed in order to refine the news content.

Nouns Prepositions

Given a token, we identify the word as a Noun word by looking at its part of speech tag given by the tokeniser. If the majority sense of that word is Noun, we discard the word. Noun words dont carry sentiment and thus are of no use in our experiments. In the same way it goes for prepositions because it is also a part of parts of speech in English dictionary too.

Stop-word-Removal

Stop words play a negative role in the task of sentiment classification. Stop words occur in both positive and negative training set, thus adding more ambiguity in the model formation. And also, stop words don't carry any sentiment information and thus are of no use to us. We create a list of stop words like he, she, at, on, a, the, etc. and ignore them while scoring the sentiment.

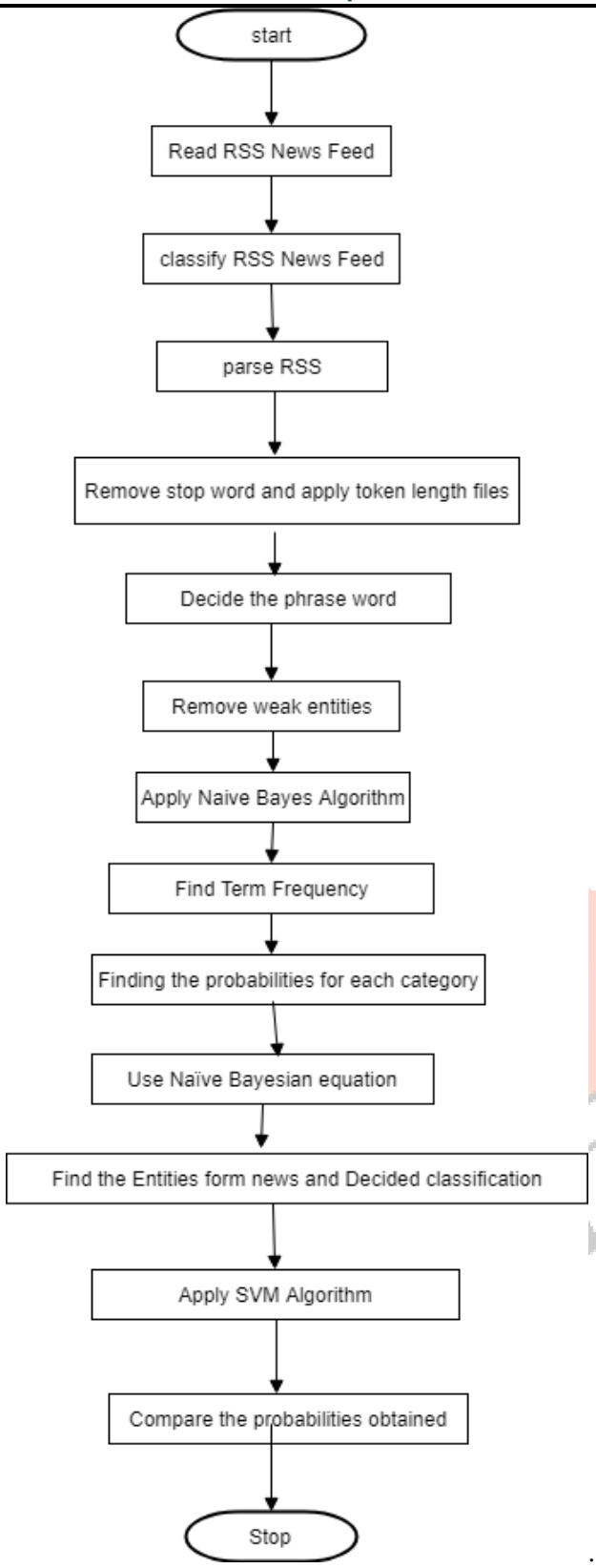

Figure.1 Showing pre-processing

II.PROPOSED SYSTEM

Nowadays, stock analysts can share their views through news document and social networking sites like twitter [3], Facebook etc. The behavior of investors is greatly affected by the sentiment of these mass media. Information is collected from RSS News Feeds. Really Simple Syndication (RSS) is a format for delivering regularly changing Web content. Many news-related sites, Weblogs and other online publishers syndicate their content as an RSS Feed to whoever wants it. It is an XML document that facilitates content syndication. A RSS is reliable way to have the web content delivered to Internet since the data is small and fast-loading, it can be used with services like cell phones or PDA's, voice mails, and email [4]. Unlike email an RSS feed is zero maintenance, the messages will never get blacklisted or filtered. With RSS, users can separate wanted information from unwanted information. RSS documents use a self-describing and simple syntax. Generally, RSS news feed contains author, title, and date information in addition to link and description. This is captured by analyses the RSS news feeds. If sentiments are correctly categorized and their polarity is correctly determined they can be helpful in enhancing a company's performance and making its investors happy. This research paper investigates the public sentiment, as expressed in large scale collections of RSS news feeds collected from stock related websites can indeed be used to predict the stock market.

A new novel approach is proposed to predict buy or sell signal to the investors in the stock market. This paper proposed a forecasting method by combining the stock related tweets and RSS news feeds with Sensex points. Based on the combined result of opinion of sentences collected from tweets, feeds and Sensex points of various stock related indicators the investors buy or sell their products. This paper explores a generic stock pre-price prediction framework and considers textual documents as inputs and generates predicted price movements as outputs. For stock price prediction, a sentence level summarization model is applied to daily full-length news article [1]. In order to obtain more accuracy, this work needs to focus on public opinion data selection.

Sentiment Analysis	Subjectivity Analysis
Positive	Subjective
Negative	Subjective
Neutral	Objective

Fig.2. Showing Analysis process

Words are classified into positive, negative and neutral. Polarity scores are calculated accordingly.

Positive	Neutral	Negative
joy	Is	Disgusting
happy	And	Sad
sweet	They	Unpleasant

Fig.3. Showing how the Classification happens

Based on the polarity scores, decision is taken.

III.ALGORITHM

We use Naïve Bayes Classifier to perform sentimental analysis on RSS News Feeds.

Naïve Bayes Classifier:

Bayesian network classifiers are a popular supervised classification paradigm. A well-known Bayesian network classifier is the Naïve Bayes' classifier is a probabilistic classifier based on the Bayes' theorem, considering Naïve (Strong) independence assumption. It was introduced under a different name into the text retrieval community and remains a popular (baseline) method for text categorizing, the problem of judging documents as belonging to one category or the other with word frequencies as the feature. An advantage of Naïve Bayes' is that it only requires a small amount of training data to estimate the parameters necessary for classification. Abstractly, Naïve Bayes' is a conditional probability model. Despite its simplicity and strong assumptions, the naïve Bayes' classifier has been proven to work satisfactorily in many domains. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined.

In Naïve Bayes' technique [9], the basic idea is to find the probabilities of categories given a text document by using the joint probabilities of words and categories. It is based on the assumption of word independence. The starting point is the Bayes' theorem for conditional probability [8], stating that, for a given data point x and class C:

$P(C / x) = P(x/C)/P(x)$ (1)

Furthermore, by making the assumption that for a data point $x= \{x1,x2,...xj\}$, the probability of each of its attributes occurring in a given class is independent, we can estimate the probability of x as follows: $P(C/x)=P(C).\prod P(xi/C)$ (2)

The Naive Bayes classifier could be an easy probabilistic classifier that relies on Bayes theorem with robust and naïve independence assumptions. Despite the naïve style and simple assumptions that this system uses, Naive Bayes performs well in several advanced real-world issues. It is typically outperformed by alternative techniques like boosted trees, random forests, liquid ecstasy Entropy, Support Vector Machines and many other tools but Naive Bayes classifier is extremely efficient since it is less computationally intensive in each central processing unit and memory and it needs a little quantity of training data. Moreover, the training time with Naive Bayes is considerably less as against different ways. The Naive Bayes classifier is a general purpose; it works well for a variety of applications. It classifies knowledge in 2 steps:

- Training step exploit the training samples, the strategy estimates the parameters of a probability distribution, assuming features are conditionally independent, given the class.
- Analysis step for any unseen test sample, the analysis find the posterior probability of that training data for each class. The method then classifies the test sample according to the largest posterior probability.

Using the Naïve Bayes classifier, the probability for a text to belong to each of the training data test can be classified [6]. The class with the very best chance for the given tweet wins, the test data is pre-processed and a vector of test data is generated. This test data is then fed into Naive Bayes classifier along with the training data, now the classifier get polarity of the highest probability.

TRAINMULTINOMIALNB(A,B)

1. V←EXTRACTVOCABULARY(B)
2. N←COUNTDOCS(B)
3. For each a€ A
4. Do N_a←COUNTDOCSINCLASS(B,a)
5. Prior[a]←N_a/N
6. $text_a$←CONCATENATETEXTOFALLDOCSINCLASS
7. for each t€V
8. do t_a←COUNTTOKENSOFTERM($text_a$,t)
9. for each t€V
10. docondprob[t][a]←(T_a+1)

APPLYMULTINOMIALNB(A,V, prior, condprob, b)

1. W ← EXTRACTTOKENSFROMDOC(V, b)
2. for each a € A
3. do score[a] ← log prior[a]
4. for each t € W
5. do score[a]+=log condprob [t][a]
6. return arg $max_{a€A}$ score[a]

Algorithm	ACC(%)	MCC
Naïve Bayes	76.30	0.46
Decision Tree	73.82	0.41
SVM	96.74	0.92

Fig.4. Showing accuracy % and Mathews Correlation Coefficient for Naïve Bayes, SVM, Decision tree algorithms

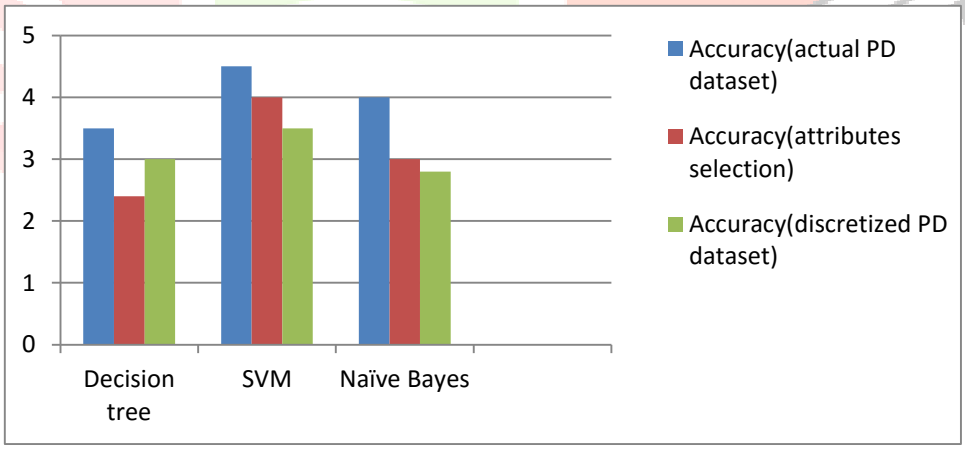

Fig.5. Showing Accuracy in various situations

SVM

Support vector machines are universal learners. Remarkable property of SVM is that their ability to learn can be independent of dimensionality of feature space. SVM measures the complexity of Hypothesis [7] based on margin that separates the plane and not number of features.

SVM learning Algorithms for Text Categorization

SVM has defined input and output format. Input is a vector space and output is 0 or 1 (positive/negative). Text document in original form are not suitable for learning. They are transformed into format which matches into input of machine learning algorithm input. For this preprocessing on text documents is carried out. Then we carry out transformation. Each word will correspond to one dimension and identical words to same dimension. As mentioned before we will see TF-IDF for this purpose. Now a machine learning algorithm is used for learning how to classify documents, i.e. creating a model for input-output mappings. SVM has been proved one of the powerful learning algorithm for text categorization.

SVM can be performed in various kernels. They are as follows:

1. Polynomial kernel
2. Gaussian kernel
3. Gaussian Radial Basis Function(RBF)
4. Laplace RBF Kernel
5. Hyperbolic Tangent Kernel
6. Sigmoid Kernel
7. Bessel Function of the First Kind Kernel
8. ANOVA radial basis Kernel
9. Linear Splines Kernel in One-dimension

Polynomial Kernel:
Equation is:
$K(x_i, x_j) = (x_{i,}.x_{j+1)}^d$

Gaussian kernel:
Equation is:
$K(x, y) = \exp(-||x-y||^2/2\sigma^2)$

Gaussian Radial Basis Function(RBF):
Equation is:
$K(x_i, x_j) = \exp(-\Upsilon||x_i-x_j||^2)$

Laplace RBF Kernel:
Equation is:
$K(x, y) = \exp(-||x-y||/ \sigma)$

Linear Splines Kernel in One-dimension:
Equation is:
$K(x, y) = 1+xy+xy \min(x, y)-((x+y)/2)\min(x, y)^2+ (\min(x, y)^3)/3$

As Linear kernel is more comfortable, we use that one.

SVM Characteristics

1. ML algorithms typically use a vector-space (attribute-value) representation of examples, mostly the attributes correspond to words. However word-pairs or the position of a word in the text may have considerable information, and practically infinitely many features can be constructed which can enhance classification accuracy.

2. Categories are binary, but generally documents are not assigned so precisely. Often a document D is said to belong a little to category X1 and a bit to category X2, but it does not fit well into any of the two. It probably would require a new category, as it is not similar to any of the documents seen before.

3. Number of words increase if we increase the number of documents. Heap's law [11] describes how the number of distinct words increases if number of document increases.

4. Representations use words as they are in texts. However, words may have different meanings, and different words may have the same meaning. The proper meaning of a word can be determined by its context i.e. each word influences the meaning of its context. However, the usual (computationally practical) representation neglects the order of the words. Task of SVM [12] is to learn and generalize the input-output mapping. In case of text categorization input is set of documents and output is their respective class. Consider spam filter as example input is an email and output is 0 or 1(either spam or no spam).
Comparing Naïve Bayes algorithm and SVM, SVM gives more accurate results.

Method	Accuracy
Baseline	73.65
Naïve Bayes	74.56
SVM	76.68
Maximum Entropy	74.93

Fig.6. Showing Accuracy of various methods

Accuracy

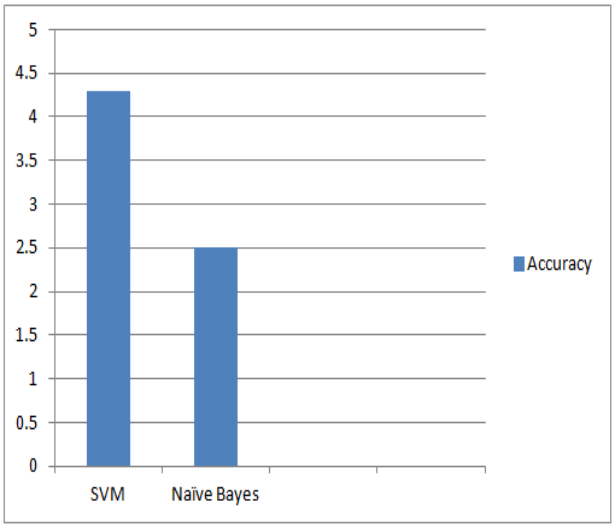

Fig.7. Algorithms showing Accuracy

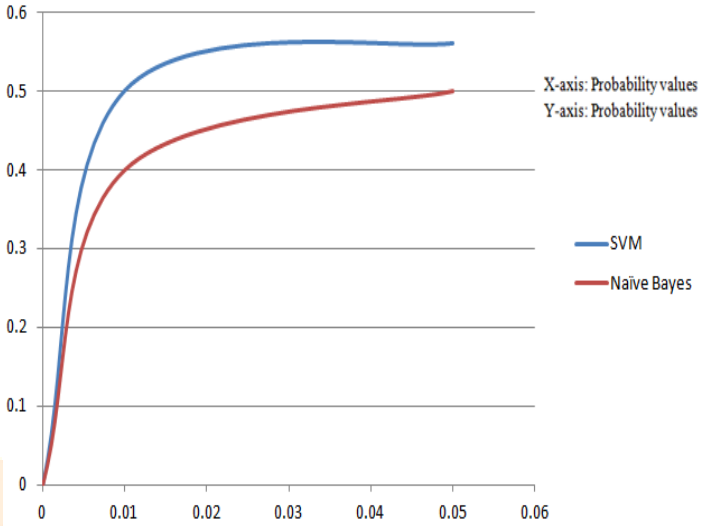

Fig.8. Graph showing the results of Naïve Bayes and SVM

These are the figures showing the accuracy of the algorithms

company	Sentimental analysis on RSS Feeds+ Moving Average Stock level indicator	Sentimental analysis on RSS Feeds+ Stock level indicator
Total Instance	650	650
Correctly Classified	530	534
Precision%	79.18%	80.11%

Fig.9.Sentiment Analysis on RSS News Feed with Precision values

SVM		Predicted		
Observed		no	yes	%
	no	91	7	91.17
	yes	9	97	88.23
				89.70
Naïve Bayes		Predicted		
Observed		no	yes	%
	no	95	6	88.23
	Yes	7	88	85.29

				86.76

Fig.10. Comparison between Naïve Bayes and SVM

Conclusion

In the conventional stock market analysis, stock level indicators such as Moving average, Moving average convergence/ Divergence, Stochastic RSI stock level indicators are used for stock market prediction. In this work, the impact of sentiments from RSS news feeds is analyzed. It is found that the sentiments from social media along with stock level indicators enhance the quality of prediction. This paper builds a predictive model to predict sentiment around stock news. First the relevant real time RSS stock news has been filtered and then they have been analyzed to predict the sentiments score values whether it is positive or negative. This is done with the help of two algorithms Naïve Bayes and SVM. So this proposed model can be a helpful tool for the investors to take the right decision regarding their stocks. Finally sentiment polarity news provides an efficient result to the stock marketers when to buy or sell their stocks. When the results obtained from two algorithms are compared, SVM gives more accurate results than that of Naïve Bayes. As the future of the system focuses on the evaluating the impact of using negation and valence shifters in addition with sentiment news, it improves the accuracy.

II. ACKNOWLEDGMENT

We like to express our gratitude for our HOD Mr R. Eshwariah for support and cooperation throughout this research paper and we would like to thank our guide Mrs. P.Jeevana Jyothi for guidance and knowledge provided by her

III. EXPERIMENTAL RESULTS

As this paper is mainly focused on Naïve Bayes and SVM algorithms, the results obtained are more accurate comparing to that of other algorithms and it is shown in the Fig.6. Compared with Naïve Bayes, SVM has more accurate results. First of all how the datasets are classified into positive or negative is shown in Fig.2. How the words in the datasets are classified into positive, negative is shown in Fig.3.
Here we considered some 10-20 datasets and classified those using Naïve Bayes and SVM and the results are as shown in Fig.10. How the pre-processing is done can be seen in Fig.1. The accuracy and precision values can be seen in Fig.8. How Naïve Bayes, SVM, Decision tree results are obtained is shown in Fig.4. The values predicted through sentiment analysis can be seen in Fig.9. The kernels used in SVM are also provided along with formulae. The results obtained using Naïve Bayes and SVM are formulated into a graph and it is shown in the Fig.7. The Analysis done using the two algorithms on RSS NEWS Feed is shown in the Fig.8. The values obtained through Sentiment analysis are more accurate in case of SVM and it can be seen in the Fig.5.

REFERENCES

[1] Jeevanandam Jotheeswaran[1], Dr. Y. S. Kumaraswamy[2,,] Opinion Mining using Decision tree based Feature Selection through Manhattan Hierarchical Cluster Measure ‘’ Journal Of Theoretical And Applied Information Technology 10th December 2013. Vol. 58 No.1

[2]X. Li, H. Xie, Y. Song, S. Zhu, Q. Li and F.L. Wang, “ Does Summarization Help Stock Prediction? A News Impact Analysis”, IEEE Intelligent Systems, Vol. 30, No. 3, pp. 26 – 34,

[3] J. Bollen, H. Mao, and X. Zeng, “Twitter Mood PredictsThe Stock Market”, Journal of Computational Science, Vol. 2,No. 1, pp. 1-8,2011.

[4] S. Bharathi and A. Geetha, “Sentiment Analysis for Effective Stock Market Prediction”, *International Journal of Intelligent Engineering and Systems,* Vol.10, No.3, pp. 146- 153, 2017.

[5]M. Usmani, S. Hasan Adil, K. Raza, and S.S.Azhar Ali, “Stock Market Prediction Using Machine Learning Techniques,” In: Proc. Of the IEEE International Conf. on Computer and Information Sciences, Vol.1, No.1, pp. 322-327,2016.

[6] Shri Bharathi1,Angelina Geetha1 and Revathi Sathiynarayanan1” Sentiment Analysis of Twitter and RSS News Feeds and Its Impact on Stock Market Prediction**”,** International Journal of Intelligent Engineering and Systems, Vol.10, No.6, 2017

[7]D.Yan,G.Zhou,X.Zhao,Y.Tian, and F. Yang,“Predicting Stock Using Micro Blog Moods”, Journal on China Communications, Vol. 13, No. 10, pp.244- 257, 2016.

[8] Miss.Kalyani D.Gaikwad1, and Prof.P.P.Rokade2” Opinion Mining using RSS Feeds and Social Media News Streams” International Conference for emerging trends in engg,technology,science and management.

[9]Z. Z. Alp and S. G.Oguduc, “Extracting Topical Information of Tweets Using Hash Tags”, In: Proc. of the IEEE International Conf. on Machine Learning and Applications, Vol.1, No.1, pp. 644-648, 2015.

[10] Q.A. Al-radaideh, A.A. Assaf, and E. Alnagi “Predicting Stock Prices Using Data Mining Techniques”, In: Proc. of the International Arab Conf. on Information Technology, Vol.4, No.3, pp. 163-191, 2013.

[11] http://www.bseindia.com/indices/indexarchivedata.aspx

[12] G. Wei, W. Zhang, and L. Zhou, "Stock Trends Prediction Combining The Public Opinion Analysis", In: Proc. of the IEEE International Conf. On Logistics, Informatics and Service Sciences, Spain, Vol.1, No.1, pp. 1-6, 2015

[13] J. Wiebe and E. Riloff, "Finding Mutual Benefit Between Subjectivity Analysis and Information Extraction",IEEE Transactions on Affective Computing, Vol.2, No.4, pp. 175-191, 2011.

[14] Q. Li, Y. Chen, L. L. Jiang, P.Li, and H. Chen, "A Tensor Based Information Framework For Predicting The Stock Market",ACM Transactions on Information Systems, Vol.34,No.2, pp.1-30, 2016.

[15] Y.Wang and Y.Wang, "Using Social Media Mining Technology to Assist in Price Prediction of Stock Market", In: Proc. of the IEEE International Conf. on Big Data Analysis,Vol.1, No.1, pp. 48-51, 2016.

[16] S.V.S. Bharathi and A. Geetha, "Sentiment Analysis for Online Stock Market News using RSS Feeds", In: Proc. of the International Conf. on Computer Science and Information Technology, Vol.1, No.1, pp. 01-07, 2017

[17] M. Kanakaraj and R. M. R. Guddeti, "Performance Analysis of Ensemble Methods on Twitter Sentiment Analysis Using NLP Techniques", In: Proc. of the IEEE International Conf. on Semantic Computing, Vol.1, No.1, pp.169-170,2015.