IJCRT Publish Paper record is available at DOI: http://doi.one/10.1727/IJCRT.17183

# Parallelism: A New Approach in Prediction System

**Kajal Borole**[1]
*Computer Department, SSBT's COET, Jalgaon*

**Satpal D Rajput**[2]
*Computer Department, SSBT's COET, Jalgaon*

*Abstract*— **Now-a -days, people have a very free and convenient communication environment of internet where they show active participation to demonstrate what they actually feel about a particular event. May it be a poll result, incident, news, political issues, government schemes or any government decision for the citizens. People demonstrate their views by reacting through various parameters available on the social media. Some make a tweet using hash tags while some other prefer giving their views using statements, updating status or even updating their profiles in order to support a particular cause. But what if the predictors already know how the public is going to react on a particular event or news. For this prediction system help us out, in which there is analysis of the responses from public and then predict the future circumstances. Many applications are developed by using prediction systems. Some example are cancer detection, election poll results and many more. The day-by-day growing data can compromise the performance of the prediction system, because its obvious that growing network of data will require more storage and the system will also consume more time for its processing. In practical applications, maintenance of network storage and calculations will be costly with increasing number of nodes in the network. For this effective strategy for reducing processing time must be introduced. To reduce this processing time introducing parallelism concept can help.**

*Keywords*— **Sentiment analysis, partitioning dataset, parallel processing, combining results.**

## I. INTRODUCTION

A prediction or forecast, is a statement about an uncertain event. It is often, but not always, based upon experience or knowledge. There is no universal agreement about the exact difference between the two terms, different authors and disciplines describe different connotations. Although guaranteed accurate information about the future is in many cases impossible, prediction can be useful to assist in making plans about possible developments. People are becoming increasingly enthusiastic about interacting, sharing, and collaborating through online collaborative media. In recent years, this collective intelligence has spread to many different areas, with particular focus on fields related to everyday life such as commerce, tourism, education, and health, causing the size of the Social Web to expand exponentially. The distillation of knowledge from such a large amount of unstructured information, however, is an extremely difficult task, as the contents of today's Web are perfectly suitable for human consumption, but remain hardly understandable to machines. Big social data analysis grows out of this need and combines multiple disciplines such as social network analysis, multimedia management, social media analytics, trend discovery, and opinion mining. For example, studying the evolution of a social network merely as a graph is very limited as it does not take into account the information flowing between network nodes. Similarly, processing social interaction contents between network members without taking into account connections between them is limited by the fact that information flows cannot be properly weighted. Big social data analysis, instead, aims to study large-scale Web phenomena such as social networks from a holistic point of view, i.e., by concurrently taking into account all the socio-technical aspects involved in their dynamic evolution. To improve the performance of prediction system such that it will be independent or least affected by the growing data network over time.

## II. LITERATURE SURVEY

According to [1] Due to the rapid development of Web, large numbers of documents assigned by readers' emotions have been generated through new portals. Comparing to the previous studies which focused on author's perspective, our research focuses on readers' emotions invoked by news articles. The research provides meaningful assistance in social media application such as sentiment retrieval, opinion summarization and election prediction. Here, the readers' emotion of news based on the social opinion network are predicted. More specifically, the opinion network based on the semantic distance is constructed. The communities in the news network indicate specific events which are related to the emotions. Therefore, the opinion network serves as the lexicon between events and corresponding emotions. Leveraging the neighbor relationship in network to predict readers' emotions is done. As a result, the methods obtain better result than the state-of-the-art methods. Moreover, we developed a growing strategy to prune the network for practical application. The experiment verifies the rationality of the reduction for application.

According to [2] With the advent to social media the number of reviews for any particular product is in millions, as there exist thousand of websites where that particular product exists. As the numbers of reviews are very high the user ends up spending a lot of time for searching best product based on the experiences shared by review writers. Here it is presented as a sentiment based rating approach for food recipes which sorts food recipes present on various website on the basis of sentiments of review writers.

The results are shown with the help of a mobile application: Foodoholic. The output of the application is an ordered list of recipes with user input as their core ingredient.

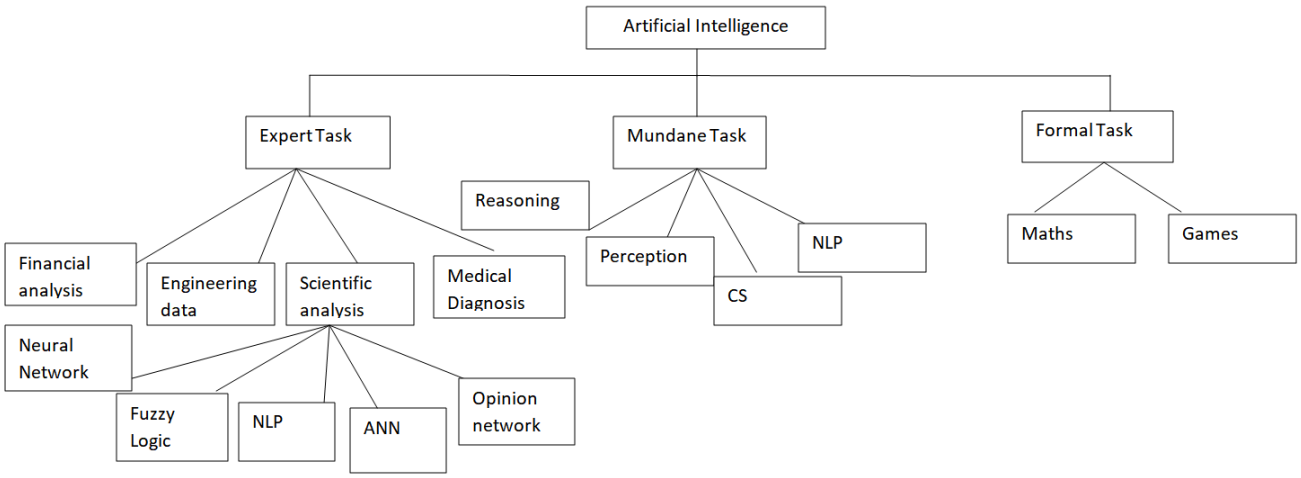


Fig.1 : Tree structure for Literature Survey

According to [3] Sentimental Analysis is one of the most popular technique which is widely been used in every industry. Extraction of sentiments from user's comments is used in detecting the user view for a particular company. Sentimental Analysis can help in predicting the mood of people which affects the stock prices and thus can help in prediction of actual prices. Here sentimental analysis is performed on the data extracted from Twitter and Stock Twits. The data is analyzed to compute the mood of user's comment. These comments are categorized into four category which are happy, up, down and rejected. The polarity index along with market data is supplied to an artificial neural network to predict the results.

According to [4] The customer review is important to improve service for company, which have both close opinion and open opinion. The open opinion means the comment as text which shows emotion and comment directly from customer. However, the company has many contents or group to evaluation themselves by rating and total rating for a type of services which there are many customer who needs to review. The problem is some customers given rating contrast with their comments. The other reviewers must read many comments and comprehensive the comments that are different from the rating. Therefore, here the proposal is to analysis and prediction rating from customer reviews who commented as open opinion using probability's classifier model. The classifier models are used case study of customer review's hotel in open comments for training data to classify comments as positive or negative called opinion mining. In addition, this classifier model has calculated probability that shows value of trend to give the rating using naive bayes techniques, which gives correctly classifier as compared with decision tree Techniques.

According to [5] Social Networking sites are the resources which contains huge data. For example, Twitter produces millions of bytes of the data. These data can be used for business or social purpose. Analyzing data from these social networking web sites is one of the new buzzword for many business strategies. Election campaigns, World health issues, Technical concepts, inventions, Entertainment, Natural resources can be effectively handle by using sentimental analysis. Our proposed work evaluates sentimental analysis of twitter data using StandfordNLP Libraries implemented in SaaS (cloud) which will handle all current affairs in the world. Cloud implementation will give process efficiency, result growth and improvement in time to market.

According to [6] Emotion Generation and Summarization form Affective Text deals with new aspect for categorizing the document based on the emotions such as Empathy, Touched, Boredom, Warmness, Amusement and Surprise. In order to predict the emotion contained in content a proposed model i.e. Emotion Topic Model is used. Using this it first generates a latent topic from emotions, followed by generating affective terms from each topic. First it separates emotion and word document and derived probabilities for it. The model which is proposed will utilize the complementary advantages of both emotion-term model and topic model. Emotion-topic model allows associating the terms i.e. words and emotions via topics which is more flexible. For classification, use of Naive Bayesian algorithm and Iteration based Nearest Neighbor Algorithm which will predict emotion accurately is done. For each emotion, it will display emoticon and songs recommendation will be available for user. So that in future user can upload and enjoy their own choice of song based on their emotion which is detected from text.

According to [7] proposing a 'Sentiment Analysis as a Service' (SAaaS) framework that abstracts sentiments from social information services, analyses and transforms into useful information help them to make their system different and robust. They proposed a dynamic service composition mechanism for sentiment analysis based on the social information service classification. Also proposed a new quality model to assess the quality of social information services. Use of social media based public health

surveillance as a motivating scenario is also discussed. In particular, focus is on the spatio-temporal properties of social media users' sentiments to identify the locations of disease out breaks. Experiments are conducted on the real-world datasets. Analytical results preliminarily show the performance of proposed approach.

According to [8] With rapid expansion of the Internet and increasing amount of time users spend online, the Internet evolves from entertainment environment towards highly dynamic and flexible business medium. Online advertisement has become one of the most successful business model for Internet environment. There are two major types of online advertisement: sponsored search and contextual display advertisement.The authors dedicated this on contextual display advertisement. Generally, contextual advertisement implementations based on topical or keyword-based relevance approach. This study addresses the mechanism of advanced contextual advertisement based on opinion about specific topic within content of webpage. Use of Natural Language Processing and Sentiment Analysis aims to determine the writer's attitude towards particular topic as: positive, negative, or neutral. This approach helps to develop an advertisement system that is more content-sensitive and consequently has higher ROI of marketing.

According to [9] Sentimental polarity detection has long been a hot task in natural language processing since its applications range from product feedback analysis to user statement understanding. Recently a lot of machine learning approaches have been proposed in the literature, e.g., SVM, Naive Bayes, recursive neural network, auto-encoders and etc. Among these different models, Convolutional Neural Network (CNN) architecture have also demonstrated profound efficiency in NLP tasks including sentiment classification. In CNN, the width of convolutional filter functions alike number N in N-grams model. Thus, different filter lengths may influence the performance of CNN classifier. Here, study the possibility of leveraging the contribution of different filter lengths and grasp their potential in the final polarity of the sentence is discussed. Then use of Adaboost to combine different classifiers with respective filter sizes is used. The experimental study on commonly used datasets has shown its potential in identifying the different roles of specific N-grams in a sentence respectively and merging their contribution in a weighted classifier.

According to [10] at present, most of the automobile engine temperature warning system are using the indicator light and sound alert to remind the drivers when the engine overheat happens, and there is no sufficient data to provide the drivers for further advices or solutions and without a database to record, that would increase the tension to cause the error operation of the drivers. Here, the temperature prediction system for the automobile engine that applies the Fuzzy-Algorithm is developed. Wherein the engine temperature, engine temperature difference (within 10 seconds) and engine revolution are the input reference factors. According to the temperature at that time as a reference to determine whether the difference between revolution and temperature variation could be dangerous for the engine. This system can indicate alert, warning dialog, automobiles status and provide a proposal for improvement on an interface of smart mobile devices as soon as the engine temperature is overheat, that reduces operator error and improves the quality of road safety.

From the above literature survey, various prediction system used various approaches and methods to predict results. But the problem remains same. The problem of improving the system is not satisfactorily discussed so far. As the data will increase in future, the data handling, processing time can be affected. To solve this a complete novel (new) approach of parallelism can help. Lets see the existing system.

## III. EXISITING SYTEM

The authors [1] conducted the prediction experiment to compare with the state of the art methods. And the main focus of the existing system was on improving the accuracy of the model. For this, based on the training data, they constructed the online news growing network. For the testing data, they added the news node into network chronologically. Then, they predicted the news emotion by their model(Social Opinion Model).Firstly, they evaluated the models presented above and discuss the properties of models. In order to ensure the rationality of the experimental results, they utilize the training data for network growing only without pruning. In prediction process, they keept the network constant to ensure the fairness of the results. The parameters are constant for the optimal value of the training network. The parameters do not change during the prediction process. Then, to conduct a comprehensive comparison with various models, they compare the existing supervised unigram model(SWAT), emotion-term(ET), emotion-topic model(ETM), affective topic model (ATM), multi-label supervised topic model (MSTM), sentiment latent topic model (SLTM), Contextual Sentiment Topic Model(CSTM) and deep learning approaches Convolutional Neural Network (CNN),CNN-SVM. For the problem of parameter tuning in topic-level models, they experimented separately under different topic numbers according to the literature, and averaged the result as the final result. For CNN and CNN-SVM, we use the CNN network structure. The results from the above existing method do not take into account the network reduction. In practical applications, maintenance of opinion network storage and calculation will be costly with the increasing number of nodes. Moreover, the social opinion towards specific news will change over time. Outdated news will affect the prediction results. More precisely, we do not need to remember the news 10 years ago [1]. To develop a system which will be least affected by the increasing data size overtime and which can process the large dataset in limited time is a challenge. To develop such a system, a novel approach of using parallelism in sentiment analysis can come to rescue. Let us see this in the proposed solution for the above problem arrived.

IJCRT Publish Paper record is available at DOI: http://doi.one/10.1727/IJCRT.17183

## IV. PROPOSED SYSTEM

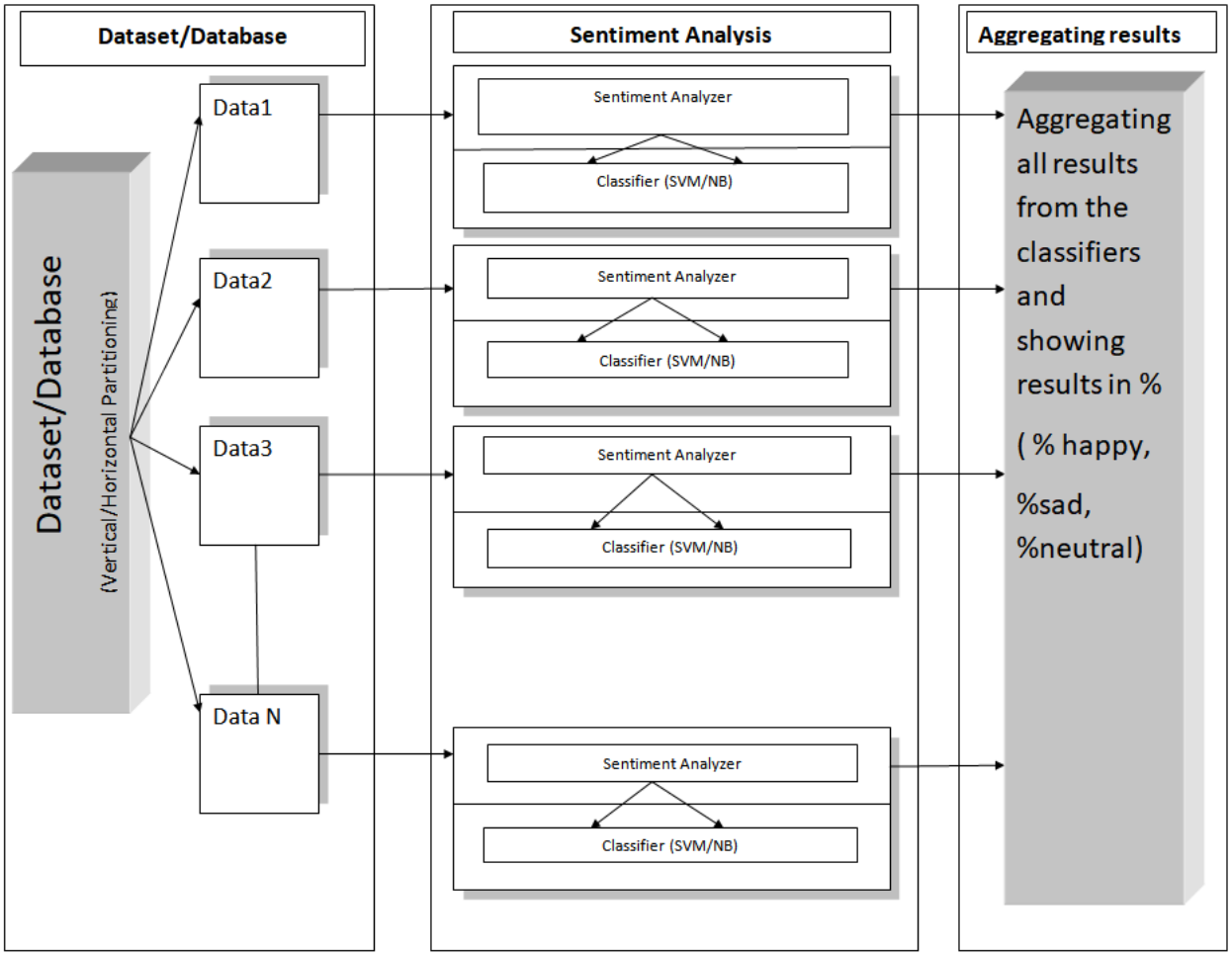


Fig. 2 : Proposed architecture for the achieving parallelism in sentiment analysis.

### *A. Partitioning Dataset/Database.*

Horizontal partitioning divides a table into multiple tables. Each table then contains the same number of columns, but fewer rows. For example, a table that contains 1 billion rows could be partitioned horizontally into 12 tables, with each smaller table representing one month of data for a specific year. Any queries requiring data for a specific month only reference the appropriate table.

Determining how to partition the tables horizontally depends on how data is analyzed. You should partition the tables so that queries reference as few tables as possible. Otherwise, excessive UNION queries, used to merge the tables logically at query time, can affect performance. For more information about querying horizontally partitioned tables. Partitioning data horizontally based on age and use is common. For example, a table may contain data for the last five years, but only data from the current year is regularly accessed. In this case, you may consider partitioning the data into five tables, with each table containing data from only one year.

Vertical partitioning divides a table into multiple tables that contain fewer columns. The two types of vertical partitioning are normalization and row splitting: Normalization is the standard database process of removing redundant columns from a table and putting them in secondary tables that are linked to the primary table by primary key and foreign key relationships.Row splitting divides the original table vertically into tables with fewer columns. Each logical row in a split table matches the same logical row in the other tables as identified by a UNIQUE KEY column that is identical in all of the partitioned tables. For example, joining the row with ID 712 from each split table re-creates the original row.

Like horizontal partitioning, vertical partitioning lets queries scan less data. This increases query performance. For example, a table that contains seven columns of which only the first four are generally referenced may benefit from splitting the last three columns into a separate table.

Vertical partitioning should be considered carefully, because analyzing data from multiple partitions requires queries that join the tables. Vertical partitioning also could affect performance if partitions are very large.

### *B. Parallel Sentiment Analysis of each partitioned dataset/database.*

Parallel Processing Systems are designed to speed up the execution of programs by dividing the program into multiple fragments and processing these fragments simultaneously. Such systems are multiprocessor systems also known as tightly coupled systems.

Parallel systems deal with the simultaneous use of multiple computer resources that can include a single computer with multiple processors, a number of computers connected by a network to form a parallel processing cluster or a combination of both. Parallelism can be attained using multithreading concept which have the capability of executing multiple thread at a same time. Each thread carrying the sentiment analysis module. Multithreading in java is a process of executing multiple threads simultaneously. Thread is basically a lightweight sub-process, a smallest unit of processing. Multiprocessing and multithreading, both are used to achieve multitasking. But we use multithreading than multiprocessing because threads share a common memory area. They don't allocate separate memory area so saves memory, and context-switching between the threads takes less time than process. Multithreading is mostly used in games, animation etc.

Advantages of Multithreading:

1) It doesn't block the user because threads are independent and you can perform multiple operations at same time.
2) You can perform many operations together so it saves time.
3) Threads are independent so it doesn't affect other threads if exception occur in a single thread.

The model for the sentiment analysis is further divided in two stages: 1. Sentiment analyser 2.Classifier as shown in above proposed architecture. Detail of sentiment analyser and classifier stages are given in figure3 and figure4 below :

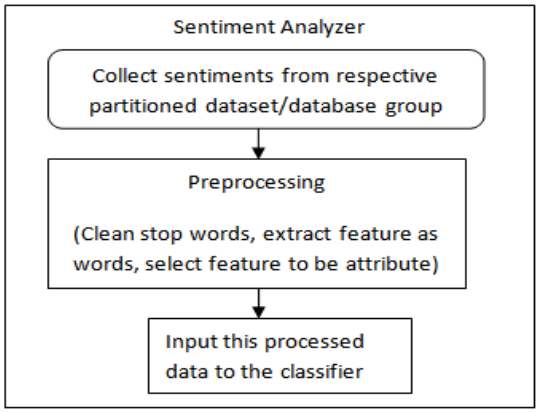


Fig.3 : Sentiment Analyzer

In Sentiment Analysis, sentiment Analyzer will take input as the one dataset from the partitioned dataset and preprocess it. There will be n no of sentiment analyser running at the same time, that is, processing at the same time. Each sentiment analyser will create a cleaned tweet sentiment words/data that will be an input to the classifier.

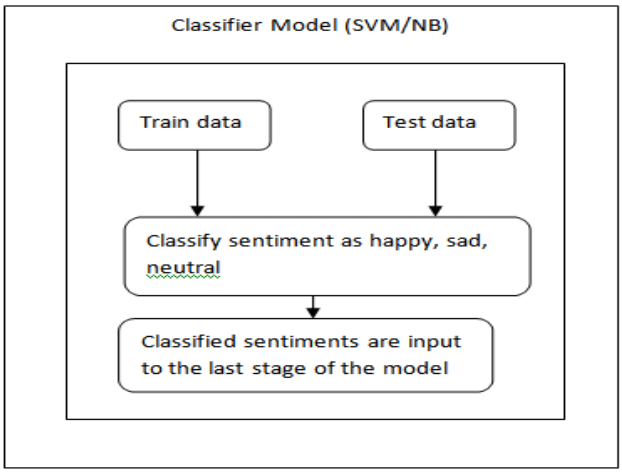


Fig. 4 : Classifier Model (SVM/NB)

Further, the when the classifier will get input from the sentiment analyser, the classifier will start its work of training and testing the data received from the sentiment analyser and will classify the data/sentiment as happy, sad or neutral. The classifier used can be any among SVM (Support Vector Machine Classifier) or NB(Naive Bayes Classifier) because so far these have the better prediction results. The output from the entire sentiment analysis module (Sentiment analyser + Claasifier) will be an input to the very final stage of the proposed architecture which is aggregating results module.

*C. Aggregating Results.*

*I. Bootstrap Aggregation (Bagging)*

Bootstrap Aggregation (or Bagging for short), is a simple and very powerful ensemble method.An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions

than any individual model. Bootstrap Aggregation is a general procedure that can be used to reduce the variance for those algorithm that have high variance. An algorithm that has high variance are decision trees, like classification and regression trees (CART). Decision trees are sensitive to the specific data on which they are trained. If the training data is changed (e.g. a tree is trained on a subset of the training data) the resulting decision tree can be quite different and in turn the predictions can be quite different. Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees. For example, let's assume we have a sample dataset of 1000 instances (x) and we are using the CART algorithm. Bagging of the CART algorithm would work as follows:

1) Create many (e.g. 100) random sub-samples of our dataset with replacement.
2) Train a CART model on each sample.
3) Given a new dataset, calculate the average prediction from each model.

For example, if we had 5 bagged decision trees that made the following class predictions for a in input sample: blue, blue, red, blue and red, we would take the most frequent class and predict blue.When bagging with decision trees, we are less concerned about individual trees overfitting the training data. For this reason and for efficiency, the individual decision trees are grown deep (e.g. few training samples at each leaf-node of the tree) and the trees are not pruned. These trees will have both high variance and low bias. These are important characterize of sub-models when combining predictions using bagging.The only parameters when bagging decision trees is the number of samples and hence the number of trees to include. This can be chosen by increasing the number of trees on run after run until the accuracy begins to stop showing improvement (e.g. on a cross validation test harness). Very large numbers of models may take a long time to prepare, but will not overfit the training data. Just like the decision trees themselves, Bagging can be used for classification and regression problems.

*II. Staking*

Stacking (sometimes called *stacked generalization*) involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all of the other algorithms are trained using the available data, then a combiner algorithm is trained to make a final prediction using all the predictions of the other algorithms as additional inputs. If an arbitrary combiner algorithm is used, then stacking can theoretically represent any of the ensemble techniques described in this article, although in practice, a single-layer logistic regression model is often used as the combiner.

Stacking typically yields performance better than any single one of the trained models. It has been successfully used on both supervised learning tasks (regression, classification and distance learning) and unsupervised learning (density estimation).It has also been used to estimate bagging's error rate. It has been reported to out-perform Bayesian model-averaging. The two top-performers in the Netflix competition utilized *blending*, which may be considered to be a form of stacking.

## V. CONCLUSION

Here, firstly, we have taken dataset of the twitter showing text emotions on a particular event. We have tired to divide the dataset vertically into n parts. Then, each partition of the dataset is given as a input to the sentiment analysis module in our architecture module, where all the steps included in sentiment analysis is carried out by sentiment analyser and the classifier module. The output from the sentiment analysis module is given as a input to the aggregation result module, where the results from all the classifier working in parallel(using multithreading concept) are collected and appropriate averaging method(discussed in aggregation module) can be used to predict accurate results within very less time as compared to the serial sentiment analysis model. Thus parallelism can be attained, and the problem of ever growing dataset or database can be managed using this approach. We hope that this review about the existing and proposed system can help us develop such a new approach which will yield good and faster results in improving prediction system.

## ACKNOWLEDGMENT

This is a great pleasure and immense satisfaction to express my deepest sense of gratitude and thanks to everyone who has directly or indirectly helped me in completing my paper work successfully. I express my gratitude towards guide Satpal D. Rajput who guided and encouraged me in completing the work in scheduled time. I am very thankful to acknowledgement to Prof. Dr. Girish K. Patnaik (HOD, Computer Engineering). I present my sincere thanks to Prof.Dr.K S.Wani, Principal for moral support and providing excellent infrastructure in carrying out the paper work. No words are sufficient to express my gratitude to my family for their unwavering encouragement. I also thank all friends for being a constant source of my support.

## REFERENCES

[1] Xintong Li, Qinke Peng, Zhi Sun, Ling Chai, and Ying Wang,*Predicting Social Emotions from Readers' Perspective*, IEEE Transaction 2017.
[2] Anshuman, Shivani Rao, Misha Kakkar *A Rating Approach based on Sentiment Analysis*, Amity University, Uttar Pradesh, IEEE 2017.
[3] Sunil Kumar Khatri, Ayush Srivastava, *Using Sentimental Analysis in Prediction of Stock Market Investment*, Amity Amity University, Noida, India, IEEE 2016.
[4] Wararat Songpan, *The Analysis and Prediction of Customer Review Rating Using Opinion Mining*, Department of Computer Science, Faculty of Science, Khon Kaen University Khon Kaen, Thailand, IEEE 2017
[5] Hase Sudeep , Kisan Hase ,Anand Kisan ,Aher Priyanka Suresh ,*Collective Intelligence & Sentimental Analysis of Twitter Data By Using StandfordNLP Libraries with Software as a Service (SaaS)*, Department of IT AVCOE, Sangamner, M.S. (India) , IEEE 2016.

[6] Sayalee S. Raut, Kavita P. Shirsat , *Implementation of Emotion Generation and Summarization form Affective Text* ,GRD Journals.
[7] Kashif Ali, Hai Dong, Athman Bouguettaya, Abdelkarim Erradi, Rachid Hadjidj, *Sentiment Analysis as a Service: A social media based sentiment analysis framework* , School of Science, Royal Melbourne Institute of Technology, Australia, IEEE 2017 International Conference.
[8] Abzetdin Z. Adamov, Eshref Adal , *Opinion Mining and Sentiment Analysis for Contextual online-Advertisement*, Qafqaz University Baku, Azerbaijan .
[9] Yazhi Gao, Wenge Rong, Yikang Shen, Zhang Xiong,*Convolutional Neural Network Based Sentiment Analysis using Adaboost Combination*, IEEE 2016 .
[10] Jui-Chuan Cheng, Te-Jen Su, Yu-Ming Cao, Chien-Yuan Pan ,*The Temperature Prediction System of Automobile Engine Based on Fuzzy Algorithm*, IEEE-ICAST 2017 .