INTERNATIONAL JOURNAL OF TECHNOLOGY AND COMPUTING (IJTC)
ISSN-2455-099X,
Volume 2, Issue 4 April 2016.

A REVIEW PAPER ON DIFFERENT CLASSIFICATION TECHNIQUES USED IN NEWS SENTIMENT ANALYSIS

Prabhjot Kaur [a], Rupinder Kaur Gurm[b]

[a] MTech Student, er.prabhjotsandhu@gmail.com,
RIMT-IET Mandi Gobindgarh , Punjab, India
[b] Assistant Professor, rupindergurm@gmail.com,
RIMT-IET Mandi Gobindgarh , Punjab, India

ABSTRACT

The data mining is the interdisciplinary subfield of computer science. The goal of data mining process is to extract information from the data set and transform it into an understandable form for further use. The objective of this paper is to describe the classification techniques used in financial news sentiment analysis. Sentiment analysis is widely use to movie reviews and social media for a variety of applications, ranging from marketing to customer service. Generally, the aim of sentiment analysis is to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation. The attitude may be bad or good. Sentiment analysis of financial news deals with the identification of positive and negative news.

Keywords—DATA MINING; SENTIMENT ANALAYSIS; CLASSIFICATIONS; REFERENCES;

I. INTRODUCTION

The data mining is the interdisciplinary subfield of computer science. The goal of data mining process is to extract information from the data set and transform it into an understandable form for further use. Generally data mining is the process of analyzing the data from different perspectives and summarizing it into useful information. Data mining is mostly used by companies with a consumer focus retail, financial, communication and marketing organizations. It enables these companies to determine relationships among internal factors such as price, staff skills, and product positioning and external factors such as economic indicators, communication etc. Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining-descriptive and Classification and Prediction. In the descriptive data mining it deals with the general properties of data in database such as mining of associations, mining of clusters, and mining of correlations. The purpose of classification and prediction is to be able to use the model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. This paper is working on the classification techniques used in financial news sentiment analysis. Sentiment analysis refers to the use of natural language processing, text mining and computational linguistics to identify and extract

understandable information in source materials. Sentiment analysis is widely use to movie reviews and social media for a variety of applications, ranging from marketing to customer service. Generally, the aim of sentiment analysis is to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation. The attitude may be bad or good. Sentiment analysis can be grouped into four types such as keyword spotting, lexical affinity, statistical methods, and concept-level techniques. Keyword spotting classifies text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored. Lexical affinity not only detects obvious affect words, it also assigns arbitrary words a probable "affinity" to particular emotions. Sentiment analysis of financial news deals with the identification of positive and negative news. In the research paper news articles present great challenge that it avoids the emotions or attitudes. Financial news is a mechanism that allows people to easily buy and sell financial assets such us stocks, commodities and currencies, among others. The main stock markets such as New York Stock Exchange, NASDAQ or London Stock Exchange have been modeled in the ontology as subclasses of the Stock market class.

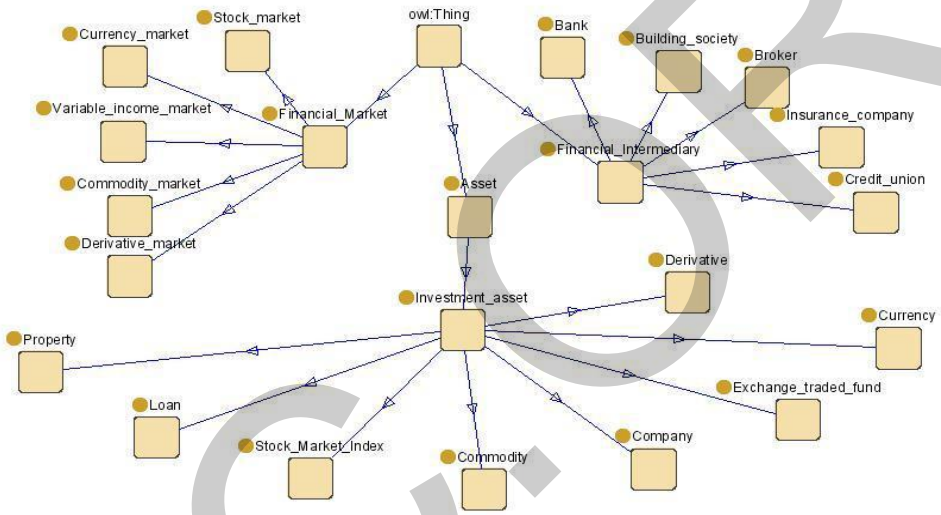


Fig. 1: Diagram of financial ontology

II. RELATED WORK

I have studied various research paper based on news sentiment analysis, in which the authors have worked on several cases of data mining such as dataset, labeling approaches of data news, feature processing which include feature extraction and feature selection, and machine learning methods for classification.

The author "Mostafa Karamibekr, Faculty of Computer Science University of New Brunswick Fredericton, NB, Canada" [1] has worked on the sentiment analysis of social issue. In this research paper the author has conducted a statistical investigation on the differences between sentiment analysis of products and social issues. To find the difference between products and social issue the author has used the different techniques such as SVM (Support Vector Machine), and Unsupervised Techniques. The unsupervised techniques are used to classify the sentiment polarity of a document. On the statistical analysis the author's research paper showed that the social issues are different from products and services because it is not easy to define

features for social issues as the case for products and services. Moreover, while in the domain of products and services, adjectives are more descriptive; in the social domains verbs are more useful to express opinions. Author has concluded that the traditional classification techniques and feature-based sentiment analysis may not be applicable for sentiment analysis of social issues.

The author "V.K. Singh, R. Piryani, A. Uddin, Department of Computer Science, South Asian University, and New Delhi, India" [2] has worked on specific features based on the sentiment analysis of movie reviews. In this paper author has used a SentiWordNet based scheme with two different linguistic feature selections comprising of adjectives, adverbs and verbs and n-gram feature extraction. He has also used SentiWordNet scheme to compute the document-level sentiment for each movie reviewed and compared the results with results obtained using Alchemy API. To find the accurate result he has used the two schemes such as SWN (AAC) and SWN (AAAVC). SWN (AAAVC) produces the most accurate results with verb score weight age factor of 30%. The SWN (AAC) method is close to the performance level of SWN (AAAVC), but it's the later method which has a marginal edge over it.

The author "Prashant Raina, School of Computer Engineering, Nanyang Technological University, Singapore" [3] has worked on Sentiment Analysis in News Articles Using Sentic Computing. In this paper the author has found the 71% result accuracy in classification with 91% precision for neutral sentences and F-measures 59%, 66% and 79% for positive, negative and neutral sentences, etc. The conclusion of this paper is that this paper is feasible to use sentic computing for fine-grained sentiment analysis in news articles.

The author "Pal-Christian and Jon Atle Gulla, Department of Computer and Information Science, Norwegian University of Science and Technology" [4] has worked on sentiment analysis of financial news. In this paper the author's purpose is to evaluate the features of financial news analysis. The conclusion of this paper is that author has found J48 classification trees to yield the highest classification performance, closely followed by Random Forrest (RF), in line studies and in opposition to the antedated conception that Support Vector Machines (SVM) is superior in this domain.

The author "Jinyan Li, Simon Fong, Yan Zhuang Department of Computer and Information Science University of Macau Taipa, Macau SAR" [5] has worked on hierarchical classification of sentiment analysis. The author has evaluated the effects of the approach in different combination of classification algorithms and filtering schemes.

The author "S Padmaja, Dept. of CSE, UCE, Osmania University, Hyderabad" [6] has compared the sentiment news articles. The Author's comparison study focused on detecting the polarity of content i.e., positive and negative effects from good or bad news for three different Indian political parties. Thus by extracting the average predicted performance author observed that the choice of certain words used in political text was influencing the Sentiments in favor of UPA which might be one of the causes for them be the winners in Elections 2009.

III. PROPOSED WORK

Objectives:

1) Collection and preprocessing of raw data for newspaper articles.
2) Filtration and feature selection using n-grams.

3) Applying hybrid classification algorithm on collected data.
4) Analyze the performance and compare it with the existing algorithm.

IV. PROBLEM FORMULATION

In the base paper, SVM algorithm is used that depends on the choice of the kernel for the classification like linear and radial basis used in paper. Also SVMs is the highly algorithmic complex. Therefore we have proposed a new classification algorithm approach by using AdaBoost algorithm for the same and also improving the performance of AdaBoost. Boosting is the machine learning method for improving the performance of any learning algorithm on the idea of creating a high accurate prediction rule by combining various weak classifiers and non appropriate rules. It was first presented by Schapire and Freud. Further research on boosting techniques introduced a new boosting algorithm called AdaBoost Algorithm.

AdaBoost Features:

1. Programming of AdaBoost is easy and it gives better and quick results.
2. AdaBoost Works fine with other different machine learning algorithms.
3. AdaBoost works well with large number of training datasets.
4. The Weak Learners cannot be too complex or too simple.

V. METHODOLOGY

1) Collection of raw data and then apply filtering techniques to make that raw data into structured format: Filtering techniques like String to Word Vector and n- gram feature selection.

2) Applying the AdaBoost algorithm on the collected data and classify the data according to the class attribute.

AdaBoost Algorithm

It assigned a weight for each leaning object. After training the previous classifier, weight of the learning objects is updated so that next classifier pay more attention to the object if it is not accurately classified by previous classifier. The assigned weight is used to vote for each classifier. If there is less error rate of classifier then more weight assigned to its vote. This training process is repeated. The weight of classifiers which voted for an object of a class is added. The class which gains higher total weight is the final class and it will introduced as the predictive class for that object.

Learning Algorithm is Decision stump

Model generation

Assign equal weight to each training instance
For *t* iterations:
    Apply learning algorithm to weighted dataset, store resulting model
    Compute model's error *e* on weighted dataset
        If *e* = 0 or *e*>= 0.5:
            Terminate model generation
    For each instance in dataset:
    If classified correctly by model:
        Multiply instance's weight by *e*/(1-*e*)
        Normalize weight of all instances

Classification

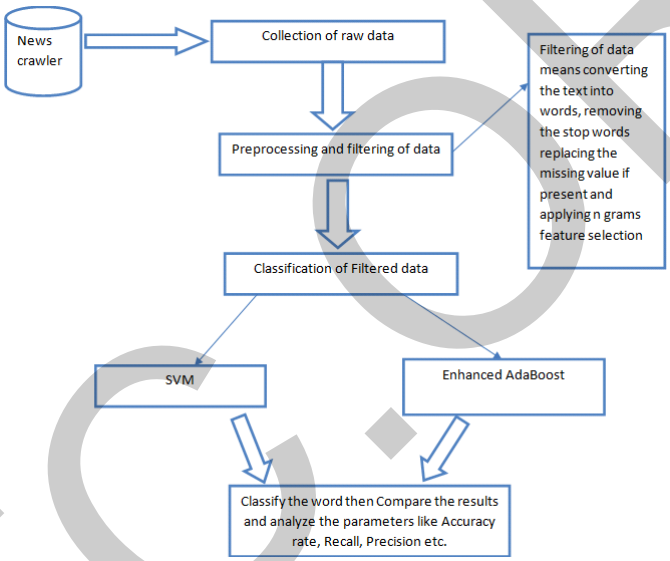
Assign weight = 0 to all classes
For each of the *t* (or less) models:
For the class this model predicts
Add –log *e*/ (1-*e*) to this class's weight
Return class with highest weight

Apply the enhanced AdaBoost algorithm for classification.

1. Replace the weak learner of AdaBoost with hybrid classifier that contains AdaBoost algorithm is hybridized on the basis of average of their probabilities.
2. Add more decision making conditions while calculating the class for model prediction i.e. on the basis of error rate adds more weight to class that will give better class for the prediction.

4) Analyze the performance parameters like FP rate, TP rate, Recall, Precision of SVM, AdaBoost and new proposed enhanced algorithm and Compare the results of three.

FLOW CHART



V TECHNIQUES USED IN DATA MINING

There are different types of data mining techniques used in sentiment analysis such as SVM (Support Vector Machine), Unsupervised, Machine Learning, RF, J48, KNN (Nearest Neighbor), Naïve Bayes and other sentiment techniques that are helpful in finding the results in research papers.

1. SVM (Support Vector Machine) :- **support vector machines** (**SVMs**, also **support vector networks**) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. An SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. When

data is not labeled, a supervised learning is not possible, and an unsupervised learning is required, that would find natural clustering of the data to groups, and map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering (SVC).

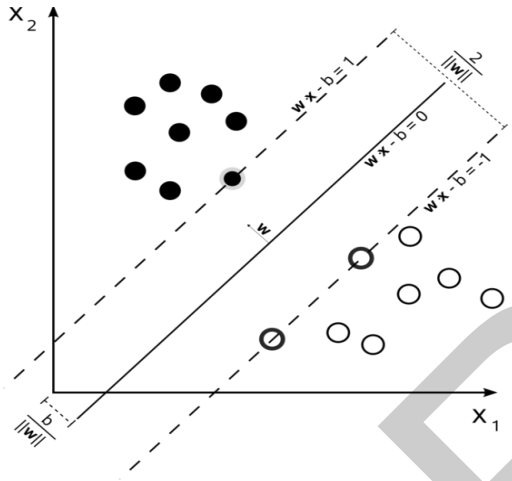


Fig. 2: SVM max hyper plane

2. Unsupervised learning: - Unsupervised learning is the machine learning task of inferring a function to describe the hidden structure of unlabeled data. Unsupervised learning includes the k-means, mixture model and hierarchal clustering. The approach of unsupervised learning is the method of moments. In the method of moments the unknown parameters in the model are related to the moments of one or more random variables, and thus, these unknown parameters can be estimated from the given moments. The moments are usually estimated from samples in an empirical way. The basic moments are type of first and second moments. For a random vector, the first order moment is the mean vector, and the second order moment is the covariance matrix covariance matrix means that the mean is zero. Higher order moments are usually represented using tensors which are the generalization of matrices to higher orders as multi-dimensional arrays.

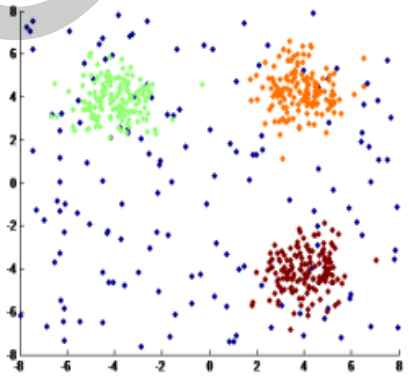


Fig. 3: Unsupervised learning

3. KNN (K-Nearest Neighbor) -The *k*-Nearest Neighbors algorithm is a non-parametric method. It is used for classification and regression. In both classification and regression, the input consists of the *k* closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors. If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors. KNN is used the method of instance based learning and lazy learning. Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor.

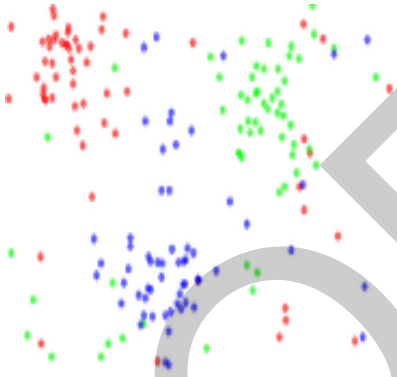


Fig. 4: KNN data Set

4. Naïve Bayes:-Naïve bayes classifier is an algorithm used to find the accuracy of news articles. Naive Bayes classifiers are highly scalable. It requires a number of parameters linear in the number of variables in a learning problem. Maximum training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. In the statistics and computer science literature, Naive Bayes models are known under a variety of names, that includes simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naive Bayes is not necessarily a Bayesian method. Author Russell and Norvig note that naive Bayes is sometimes called a Bayesian classifier. Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}.$$

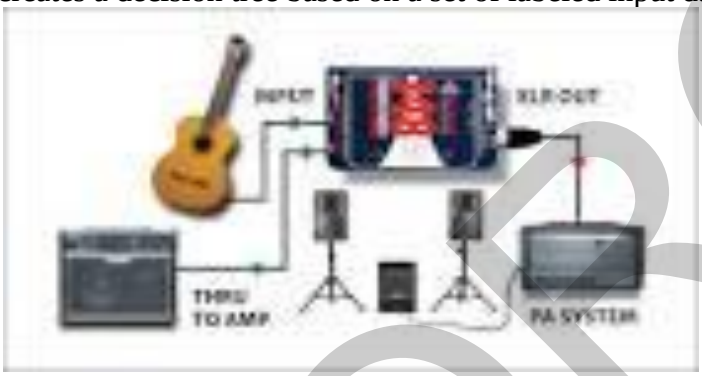
In simple the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

5. Random Forest(RF) :- Random forests is a notion of the general technique of random decision forests, that are ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode

of the classification or mean regression of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. The algorithm for inducing Breiman's random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variance. The selection of a random subset of features is an example of the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

6. J48:- J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is a program that creates a decision tree based on a set of labeled input data.



VI. CONCLUSION

I have studied many research papers. All papers are published in IEEE standards. These papers are based on sentiment analysis. In the research paper the authors have used many techniques to find the accurate result in news articles. Most of researchers have used the SVM techniques. This is called the Support vector machine. It is a based supervised learning. Others techniques used in the sentiment news articles are KNN (k-nearest Neighbor), J48 (Waka tool), RF and Naïve Bayes.

REFERENCES

[1] Mostafa Karamibekr, "Sentiment Analaysis of Social Issues," Faculty of Computer Science, University of New Brunswick Fredericton, NB, Canada, 2012 ,IEEE statndards.
[2] V.K. Singh, R. Piryani, A. Uddin, "Sentiment Analaysis of movie reviews", Department of Computer Science, South Asian University,2013 IEEE standards.
[3] Prashant Raina, "Sentiment Analysis in News Articles Using Sentic Computing," School of Computer Engineering, Nanyang Technological University ,Singapore, 2013 IEEE standards.
[4] Jon Atle Gulla, "Evaluating Features sets and classifiers for sentiment analysis of financial news ", Department of Computer and Information Science, Norwegian University of Science and Technology,2014 IEEE standards.
[5] Jinyan Li, Simon Fong, Yan Zhuang,"Hierarichal Classification in Text Mining for Sentiment Analysis", Department of Computer and Information Science,2014 IEEE standards.
[6] Padmaja," Comparing and Evaluating the Sentiment on Newspaper Articles" Department Of CSE, Hyderabad,Science and information conference 2014.