A Method Of Sentiment Analysis Using Machine Learning With Feature Engineering

Sourabh Rai[1], Prof. Shiv Tiwari[2]
[1, 2] Dept of CSE
[2]Professor, Dept of CSE
[1, 2] Shri Ram Institute of Science & Technology
Jabalpur, Madhya Pradesh, India.

Abstract- *Micro blogging websites like Twitter and Facebook, in this new era, is loaded with opinions and data. One of the most widely used micro-blogging site, Twitter, is where people share their ideas in the form of tweets and therefore it becomes one of the best sources for sentimental analysis. Opinions can be widely grouped into three categories good for positive, bad for negative and neutral and the process of analyzing differences of opinions and grouping them in all these categories is known as Sentiment Analysis. Data mining is basically used to uncover relevant information from web pages especially from the social networking sites. Merging data mining with other fields like text mining, NLP and computational intelligence we are able to classify tweets as good, bad or neutral.*

In order to improve classification results in the domain of sentiment analysis, we are using ensemble machine learning techniques for increasing the efficiency and reliability of proposed approach. For the same, we are using Linear Support Vector Machine and experimental results prove that our proposed approach is providing better classification results in terms of f-measure and accuracy in contrast to individual classifiers. We also use accuracy comparison framework for comparing algorithms based on execution time.

Keywords- Sentiment analysis, Twitter, Adjective analysis, Naive bayes, Linear SVM.

I. INTRODUCTION

Documents written in natural language constitute a major part of the artifacts produced during the software engineering life cycle [1]. According to studies [2], 85–90% of all corporate data are stored in some sort of unstructured form, mainly as text. Web news articles, abstracts of research papers, and blogs reviews are other examples of documents written in natural language which are important sources for further analysis and improved decision making. Therefore, the growth of social media and user-generated content (UGC) on the Internet provides a huge quantity of information that allows discovering the experiences, opinions, and feelings of users and customers [3]. Twitter, which is one of the most used social media, has 320 million monthly active users and it oversees 1 billion tweets everyday [4]. Since it is a rich source of real-time information, many entities such as companies, politicians, and government have demonstrated interest in knowing the opinions of people at this site.

However, manual text analysis is a time-consuming, error-prone, and costly task. In this sense, algorithms and techniques from data mining, statistics, and natural language processing (NLP) are used to extract important information from text [5]. The *text mining* term was cited for the first time by Feldman and Dagan [6] in 1995 as a *machine supported analysis of text*. Text mining is also defined as the process of knowledge discovery from text (KDT), which extracts knowledge, from unstructured datasets, commonly known as corpus. Different from data mining, text mining analyses unstructured data such as documents written in natural language from various data sources including news articles, social networks, blogs, research papers, web pages, journals, reports, software engineering artefacts, and others [7]. While data mining is largely language independent, text mining requires substantial linguistic knowledge, justifying its study together with a target language.

The internet technology has not only brought people together by connecting them on social-networks but has also played an important role in the expansion of e-commerce. Amazon, Snapdeal, Taobao, Eopinion, etc. are one of those e-commerce websites which not only sell the products online, but also provide a platform where the customers are allowed to post the reviews about the purchased products [8]. Research shows that online customer product reviews not only have significant impact on customers' online purchase decisions but are also helpful for the manufacturers to improve the product design and quality and for the online retailers to improve their services [9], [10]. Lengthy reviews make it hard for the online customers to read full reviews in order make a decision on whether to purchase the product or not. On the other hand, reading incomplete reviews might give a prejudiced view to the customers [11]. Another problem that is frequently quoted in many studies, is regarding the customer preferences for

different product features [12], [13]. This leads to a finale that a particular review, though may be descriptive but may not be helpful to a customer who is looking for the features not mentioned in that review. There are a very few online review platforms which care about organizing the reviews in manner that is feature oriented and customer friendly [14]. Many researchers are working in the field of opinion mining and sentiment analysis to extract product specific features [15].
In general, feature based opinion mining involves three subtasks viz.

(i) To correctly identify the opinionated and product specific features,
(ii) To identify the review sentences attributing positive/ negative opinions to the extracted features and
(iii) To generate a feature based summary from the information extracted.

The aim is to improve the accuracy and simplify the task of mining the opinions of customer reviews with respect to the features extracted.

1.1 Applications of Opinion Mining:

The various applications of Opinion Mining are as follows:

• It is used in E-commerce. Whenever the customer purchases any item, then it allows them to submit their opinions regarding the quality of products or services. A summary for the product and various features of the products are provided by assigning rating.
• It is used in Entertainment to determine the polarity of the movie by helping people to choose which movie or series to watch.
• Sentiment analysis can be used by policy makers who can look into the citizen's point of view towards various policies and this information can be utilized in creating better citizen friendly policies.
• It is used in Marketing, where each company enables facilities to their users to provide opinions about their products and services. It is helpful for businesses to save money and time because there is no need to conduct surveys as the feedbacks related to all the products are available on their sites.
• It is used in the field of education to help students to determine which university is good for studies.

II. RELATED WORK

2.1 Machine Learning Model

Machine Learning is the ability of machines to learn, where a machine is built up using certain algorithms through which it can take its own decisions and provide the result to the user. Basically it is considered the subfield of Artificial Intelligence. Today Machine Learning is used for complex data classification and decision making [16]. In simple terms it is the development of algorithms that enables the system to learn, and to make necessary decisions. It has strong ties to mathematical optimization that delivers methods, theory and application domain to the field and, it is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. Certain examples applications are Spam filtering, optical character recognition (OCR), Search Engines and Computer Vision.

Machine Learning methods and tasks are broadly divided into three categories as follows.

• Supervised Learning
• Un-Supervised Learning
• Reinforcement Learning

A. Supervised Learning

In this type of learning the system is provided with a sample inputs and it is mapped with the output. In this type of learning, each example is a pair consisting of an input object (basically a vector) and a desired output value (supervisory signal). A supervised learning algorithm analyses and studies the training data and produces an inferred function that can be used for mapping new examples. The optimal scenario will allow the algorithm to correctly determine the class labels for unseen instances. It is required by the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way. Approaches for Supervised Learning are Support Vector Machines, Decision Trees, etc. [17].

B. Un-Supervised Learning

In this type of learning the system is provided with some sample inputs but there is no any output present. Since there is no desired output over here categorization is done so that the algorithm differentiates correctly between the data sets. It is a task of defining a function to describe hidden structure from unlabelled data. Since samples or training sets given to the learner are unlabelled, there is no error to reward signal to evaluate a potential solution. In this way unsupervised learning differs from supervised learning and reinforcement learning. It is closely related to the problem of Density Estimation and statistics [17].

C. Reinforcement Learning

Reinforcement learning is a sub domain of machine learning inspired by behaviorist psychology, dealing with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. It is studied and used in many theories like game theory, control theory, operations research, information theory, swarm intelligence, statics and genetic algorithms [17]. This paper is majorly focusing on studding different algorithms of machine learning which would help system to accurately classify the data and would enable it for decision making in complex situations. The two main algorithms to study are decision tree algorithm and support vector machines.

2.2 Support Vector Machine

SVM is into picture since 1992, when there was a need of classification and regression tools based on some predictions. It is introduced by Vapnick, Guyon and Boser in COLT-92. For separating any data we define certain classes and depending on the complexity of the datasets we define it as the linear or nonlinear classification. SVM can just be defined as a prediction tool wherein we search for a particular line or decision boundary termed as hyperplane which easily separates out the datasets or classes, hence it avoids the extra over fit to the data. It uses hypothesis space of a linear space into a high dimensional feature space. It is also capable of classifying the nonlinear data where it uses kernel functions.
SVM for Linear Classification Support Vector Machine is used for classification and Regression. It is a novel strategy of separating the samples by just drawing a decision boundary known as hyper plane in case of linear classification. For selection of hyper plane we follow the below steps.

1) Define a function such that it will generate the required hyperplane i.e. boundary in between the different datasets.
2) Next step is to select a hyper plane and calculate its distance from both the sides of the datasets.
    a) If the distance which is calculated is maximum on both the sides as compared to the previous hyperplane then select this hyperplane as the new decision boundary.
    b) Mark the samples which are close to the hyperplane as the supporting vectors. (Helps in selection of decision boundary)
3) Repeat step 2 until best hyperplane is found.

There are 2 key implementations of SVM technique that are mathematical programming and kernel function. It finds an Optimal separates hyper plane between data point of different classes in a high dimensional space. Let's assume two classes for classification. The classes being P and N for Yn= 1,-1, and by which we can extend to K class classification by using K two class classifiers. Support vector classifier (SVC) searching hyper plane. But SVC is outlined so kernel functions are introduced in order to non line on decision surface.

Sentiment Analysis is the thorough research of how opinions and perspectives can be relate to ones emotion and attitude shows in natural language respect to an event. Recent events show that the sentiment analysis has reached up-to great achievement which can surpass the positive vs negative and deal with whole arena of behavior and emotions for different communities and topics. In the field of sentiment analysis using different techniques good amount of research has been carried out for prediction of social opinions. Pang and lee (2002) proposed the system where an opinion can be positive or negative was found out by ratio of positive words to total words. Later in 2008 the author developed methodology in which tweet outcome can be decided by term in the tweet. Jiang (2011) and Tan (2011) have applied maximum entropy (Max-Ent), Naïve Bayes (NB) and support vector machines (SVM) as supervised classifiers [18]. Chen (2011) employed the feed-forward BPN network and uses sentiment orientation to calculate the results at each neuron [19].

Malhar and Ram (2014) employed supervised machine learning techniques and artificial neural networks to classify twitter data along with case study of Presidential and Assembly elections which results SVM outperforms all other classifiers [20]. Anton and Andrey reviewed the existing techniques and developed a model for automatic sentiment analysis of twitter messages using unigram, bigram and jointly i.e. hybrid feature [21]. Pak and Paroubek (2010) perform linguistic analysis and build a sentiment classifier to determine positive, negative and neutral sentiments for a document. Tang, Tan and Cheng exchanges views on main approaches and issues to problems like word sentiment classification, opinion extraction, subjectivity classification and document sentiment classification. Sentiment classifier can be prevented from probably misguiding or irrelevant text by subjective classification. Kopel and Schler explain that it is very important to use neutral messages to get good knowledge of polarity. The authors also states that positive and negative messages alone will not give proper understanding about neutral messages.

Several methods are available in the literatures, which use base classifiers for Twitter SA. Medhat et al. [22] presents a survey of Sentiment Analysis algorithms and applications. Davidov et al. [23] and Go et al. determined sentiments by using emoticons and hashtag. Linguistic processing is not fully covered in this paper. The result shows

that ensemble classifier performed well. Rodriguez et al. used N-gram, lexicon, POS and Sentiwordnet as feature set. SVMs and Conditional Random Fields are used as base learner. Their ensemble combination of orthogonal methods leads to more accurate classifiers. Hassan et al. developed an ensemble technique which used dataset, feature set and bootstrap aggregation learners. They proposed an algorithm that would select the most appropriate classifier among all the base classifiers. Clark et al. proposed an ensemble classifier which is trained on features like lexical to determine the polarity of each individual phrase within each tweet. The sentiment of a specific phrase may not be same as the sentiment of the whole tweet.

III. PROPOSED SYSTEM

Proposed system is a two-step approach. As seen in Figure 3.1, firstly, data is collected and preprocessed in data preparation step. After data is preprocessed and corrected, Aspect & Feature Based Sentiment Analysis step starts, which is the main focus of this thesis. At the beginning of this step, the sentences extracted in the Data Preparation part and are passed to Feature based processing and Aspect Extraction module. Then, these extracted aspects and the sentences are given to Sentiment Classification module and for all sentences in the dataset the orientation of thoughts over the extracted aspects are determined by using sentiment words in these sentences.

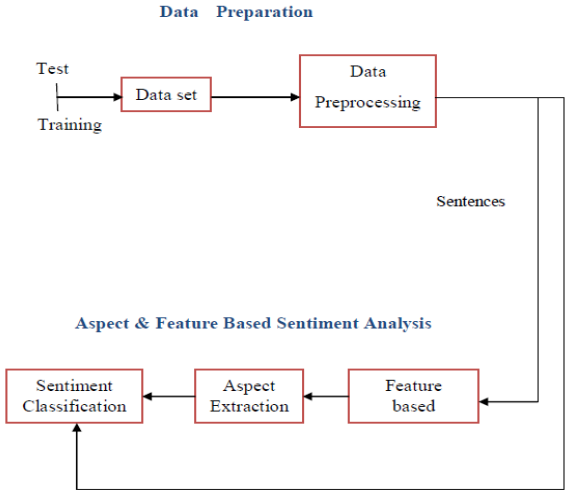

Figure 3.1: Overall Architecture of Proposed System.

Sentiment analysis is defined as determining the neutrality, positivity or negativity of a given *text* based on given *aspect term*. We are to analyze a dataset, containing both, aspect term information and the text. Our solution involves understanding various approaches to tackle the task, analyzes the text/sentences and finally builds a classifier capable of determining the sentiment of the provided text/sentence. We examine various text cleaning techniques, machine learning models and discuss their respective merits.

3.2 Detailed Description of Process:

The framework used for this analysis is depicted in below. Different processing steps had their own important role. We discussed about all steps below.

A. Data Collection and Loading

Collection of data is an important part of Sentiment Analysis. Various data Sources like Blogs, Review Sites, Online Posts & Micro Blogging like Twitter, Facebook are used for Data Collection. We used Twitter for Data Collection process.

After data is collected, it is loaded into CSV format for training and testing purpose.

B. Data Pre-Processing

The 'text' and 'aspect term' columns contain a lot of symbols which are detrimental to the task of sentiment analysis. Removal of < > (em tags), _ (underscore), "(quotes), white spaces and non-ascii codes (emoji) was essential. Additionally, we explored the conversion of emoticons, which strongly represent an emotion, into specific words to expose the underlying sentiment, but found that it simply adds noise to the corpus. Perhaps this was due to various sarcastic emoticons being used by the subjects. Some symbols are also removed for reducing noises from data. Some very rare words e.g., that just occur 10 times in 100000 words are also removed. After removing symbols, data lemmatization and stemming method is performed on the datasets. After completing the process we get a clean data for further process.

C. Feature Extraction

We initially tried several features like complete n-gram vectorization of complete text data but came down to following features, as they played most important role in determining the sentiment:

- Distance between aspect term and nearest feature.
- Class of adjective.
- Positive words count.
- Negative words count.
- N-gram vectorization of aspect +5 words left & right.

IV. RESULTS & EVALUATION

Evaluations of various algorithms according to different parameters are displayed below:

The classification performance can be evaluated in three terms: accuracy, recall and precision as defined below. Accuracy explains correctly classified instances. Precision and Recall are in weighted average for positive and negative terms.

Classifier	Accuracy (in %)	Precision	Recall
Naive Bayes	65.49	0.605	0.290
Decision Tree	65.02	0.559	0.345
Linear SVM (Proposed)	**69.90**	0.875	0.932

Table 4.1 Performance Evaluation.

V. CONCLUSION

In this paper, Naive Bayes Algorithm, DECISION TREE algorithm and proposed algorithm for sentiment classification model are presented for improving the overall accuracy of the classifier in the classification of tweets. For the same we apply preprocessing techniques so that accurate data is fed as an input to the training process, our proposed approach classify the tweets as Positive and Negative tweets which further helps in sentiment analysis and uses that sentiment analysis for further decision making. The work of proposed model has gone through preprocessing stage and classifiers learning stage. For analytical evaluation of the proposed classifier accuracy, precision and recalls are used.

The comparative results prove that proposed model improved the overall classification accuracy and precision measure of sentiment prediction as compared to traditional existing techniques for classification.

REFERENCE

[1] Witte, R., Li, Q., Zhang, Y., et al.: 'Text mining and software engineering: an integrated source code and document analysis approach', IET Softw., 2008, 2,(1), pp. 3–16.

[2] Delen, D., Cross land, M.D.: 'Seeding the survey and analysis of research literature with text mining', Expert Syst. Appl., 2008, 34, pp. 1707–1720.

[3] Marine-Roig, E., Anton Clavé, S.: 'Tourism analytics with massive user generated content: a case study of Barcelona', J. Destination Mark. Manage. 2015, 4, pp. 1–11.

[4] 'Twitter Official Webpage', 2016. Available at https://about.twitter.com/company, Accessed: March, 2016.

[5] Hotho, A., Andreas, N., Paaß, G., et al.: 'A brief survey of text mining', LDV Forum – GLDV J. Comput. Linguist. Lang. Technol., 2005, 20, pp. 1–37.

[6] Feldman, R., Dagan, I.: 'Knowledge discovery in textual databases (KDT)'. Int. Conf. Knowledge Discovery and Data Mining (KDD), 1995, pp. 112–117. Available at http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf, Accessed: March 2016.

[7] Shi, G., Kong, Y.: 'Advances in theories and applications of text mining'. Int. Conf. Information Science and Engineering (ICISE2009), 2009, pp. 4167–4170.

[8] Sivarajah, Uthayasankar, Zahir Irani, and Vishanth Weerakkody, "Evaluating The Use And Impact of Web 2.0 Technologies in Local Government," Government Information Quarterly. Elsevier, pp. 473–487, 2015.

[9] Magdalini Eirinaki, Shamita Pisal, and Japinder Singh, "Feature based opinion mining and ranking," Journal of Computer and System Sciences, vol.78, pp. 1175–1184, 2012.

[10] Kushal Bafna, and Durga Toshniwal, "Feature Based Summarization of Customers' Reviews of Online Products," in proc. 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems –KES, vol. 22, pp. 142-151,2013.

[11] Minqing Hu, and Bing Liu, "Mining and Summarizing Customer Reviews," Association for Computing Machinery -ACM, pp. 168-177, 2004.

[12] Edison Marrese-Taylor, Juan D. Velasquez, and Felipe Bravo-Marquez, "A novel deterministic approach for aspect-based opinion mining in tourism products reviews," Expert Systems with Applications, vol. 41, pp. 7764–7775, 2014.

[13] Changqin Quan, and Fuji Ren, "Unsupervised product feature extraction for feature-oriented opinion determination," Information Sciences, vol. 272, pp. 16–28, 2014.

[14] Zhijun Yan, Meiming Xing, Dongsong Zhang, and Baizhang Maa, "EXPRS: An extended pagerank method for product feature extraction nbbn from online consumer reviews," Information & Management, vol. 52, pp. 850–858, 2015.

[15] Ayoub Bagheri, Mohamad Saraee, and Franciska de Jong, "Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews," Knowledge Based Systems, vol.52, pp. 201–213, 2013.

[16] U.V Kulkarni, S.V Shinde, "Neuro –fuzzy classifier based on the Gaussian membership function", 4th ICCCNT 2013, July 4-6, 2013, Tiruchengode, India.

[17] Vikramaditya Jakkula, “Tutorial on Support Vector Machine” ,2013.
[18] L. Jiang, M. Yu, M. Zhou, X. Liu and T. Zhao, "Target dependent twitter sentiment classification", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 151--160, 2011.
[19] L. Chen, C. Liu and H. Chiu, "A neural network based approach for sentiment classification in the blogosphere", Journal of Informatics, vol. 5, no. 2, pp. 313-322, 2011.
[20] M. Anjaria and R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning", Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on, pp. 1--8, 2014.
[21] A. Barhan and A. Shakhomirov, "Methods for Sentiment Analysis of Twitter Messages", 12th Conference of FRUCT Association, 2012.
[22] Huma Parveen, Shikha Pandey,” Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm” IEEE-2016.
[23] Megha Rathi, Aditya Malik et. al.,“Sentiment Analysis of Tweets using Machine Learning Approach”, IEEE-2018.