IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SENTIMENT ANALYSIS USING MACHINE LEARNING APPROACHES OF TWITTER DATA AND SEMANTIC ANALYSIS

Md Ashique, Satyam Kumar, Swapnil Panwar, Aanchal Vij

Student, Student, Student, Assistant Professor

Computer Science & Engg

Galgotias University, Noida, India

Abstract: The widespread use of the World Wide Web has ushered in a new way for people to share their feelings. It is also a medium with a wealth of knowledge where users can see other users' opinions, which are divided into various sentiment groups and are gradually becoming a key factor in decision-making. This paper contributes to the sentiment analysis for consumer review classification, which is useful for analysing information in the form of a large number of tweets with highly unstructured views that are either positive or negative, or somewhere in between. To do so, we first pre-processed the dataset, then extracted the adjectives from it that have some context, which is known as feature extraction. vector, then added the function vector list classification algorithms that use machine learning, such as: The content function is extracted using Naive Bayes, Maximum Entropy, and SVM, as well as the Semantic Orientation based WordNet, which extracts synonyms and similarity. Finally, we evaluated the classifier's output in terms of recall, precision, and accuracy.

I. INTRODUCTION

The current research paper examines the contents on the Internet in a variety of areas that are rapidly expanding in both number and volume as sites devoted to particular types of items specialize in gathering customer feedback from different sites such as Amazon. Even though Twitter is a place where tweets express thoughts, gaining a general understanding of these unstructured data (opinions) can take a long time. These unstructured data (opinions) on a particular site are used by users, who form an impression of the goods or services and, as a result, make a decision. These opinions are then aggregated to collect input for various reasons, and sentiment analysis is used to provide valuable opinions. Sentiment analysis is a method in which a dataset of feelings, attitudes, or assessments is used to consider how an individual thinks [1]. Trying to consider the positive and negative aspects of a sentence is a challenging job. To summaries the analysis, the features used to characterize the sentences should have a very strong adjective. These contents are often written in a variety of styles that are difficult for users or businesses to deduce, making it difficult to categories them. Until purchasing a product, consumers are influenced by sentiment analysis to classify whether the information about it is satisfactory or not. Marketers and businesses use this research to gain a better understanding of their goods or services so that they can be tailored to the needs of their customers. Unsupervised and supervised machine learning techniques are the two forms of machine learning techniques commonly used for sentiment analysis [2]. Unsupervised learning does not have a category and does not have the right targets at all, so it must be clustered. The model is given the labels during the process since supervised learning is based on a labeled dataset. When these labeled datasets are encountered during decision-making, they are conditioned to deliver appropriate outputs. This research paper is focused on supervised machine learning in order to help us better understand sentiment analysis.

II. RELATED WORK

A lot of work has been performed in the area of "Sentiment analysis" by a number of researchers in recent years. In reality, work in the field began at the turn of the century. It was designed for binary classification in its early stages, which assigns thoughts or reviews to bipolar groups such as positive or negative. The paper [3] foresees an analysis. Using an unsupervised learning algorithm, the average semantic orientation of a phrase containing adjective and adverb is used to determine whether the phrase is positive or negative, resulting in a thumbs up or thumbs down evaluation. Some sentiment analyses, such as [4], are focused on a review of the product's consumer summarization method. The product function in [4] employs a latent semantic analysis (LSA)-based filtering mechanism to classify opinion terms, which are then used to pick some sentences for inclusion in a rich review summarization. A contrast of positive and negative sentences is used in Paper [5.] It collects data from the web and labels the word set manually, which takes a lot of time and effort. The author of [6] used a rule-based approach for sentiment analysis of Chinese documents based on Baseline and SVM, which extracts the overall document polarity of specific words using a sentiment word dictionary and adjusts it based on context details. The polarity of the word is determined in another work [7] by all the terms in the sentence, which can be positive or negative depending on the relevant sentence structure. Lakshmi and Edward [8] suggested pre-processing the data to increase the raw sentence's consistency structure. For sentiment analysis, they used the LSA technique and cosine similarity. For sentiment classification, Basant Agarwal et al. [9] used the term pattern system. It extracts contextual and syntactic information from the document using part of speech-based rules and dependency relationships. The author of [10] intended to present aspect-focused opinion polling based on unlabeled free form textual customer reviews that do not require customers to respond to the questions. M. Karamibekr and A.A. Ghorbani [11] suggested a method for sentiment classification of a text in the social domain based on verbs as an important opinion expression. SentiFul is a sentiment lexicon created by Paper [12], which uses and expands it by synonyms, antonyms, hyponyms, derivation ,and compounding. They suggested a method for identifying four types of affixes based on their position in sentiment features: propagation, weakening, reversing, and intensifying. These methods allocate polarity to sentiments, which aids in the expansion of the lexicon and thus improves sentiment analysis. For sentiment analysis, a lot of work has been done where researchers have explored and implemented soft-computing methods, mostly fuzzy logic and neural works. [13] and [14] are two examples of works that use fuzzy logic to solve problems. [13] makes a significant contribution by using a fuzzy domain sentiment ontology tree extraction algorithm. This algorithm creates a fuzzy domain sentiment ontology tree based on feedback, which involves extracting sentiment terms, product attributes, and feature relationships, and accurately predicting the polarity of the reviews. The authors created a fuzzy inference scheme based on membership functions in [14]. They devised and standardized the method of quantifying the strength of reviewer's opinions in the presence of adverbial modifiers by designing membership functions. They used the technique to study adverbial modifier trigram patterns.

III. OUR APPROACH

We evaluated the Twitter dataset as part of our approach. The unigram feature extraction technique is used to analyse labelled datasets. We used a system in which a preprocessor is applied to raw sentences to make them more understandable. Furthermore, various machine learning techniques train the dataset with feature vectors, and then semantic analysis provides a broad collection of synonyms and similarity, which determines the content's polarity. The approach has been outlined in detail in the following subsections, and a block diagram of the approach is graphically depicted in Fig

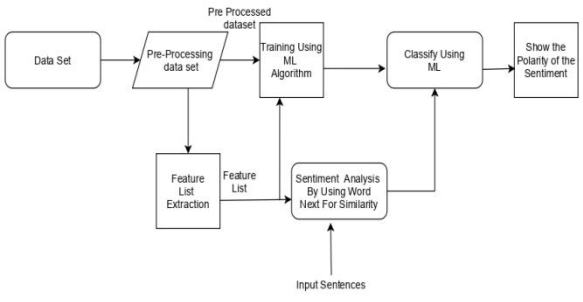

Fig.1. Diagram of the Approach to Problem

1. Pre-processing of the datasets
   Individuals express themselves in a variety of ways in the tweets, which contain a lot of opinions about the results. The dataset used in this study has already been labeled. The polarity of a labeled dataset is negative and positive, making data analysis easy. Inconsistency and redundancy are common problems with polarized raw data. The quality of the data has an impact on the performance, so raw data is pre-processed to enhance the quality. It is concerned with data preparation that eliminates repetitive words and punctuation and increases data quality. After preprocessing, the phrase "the tajmahl is Beauutifull #" becomes "tajmahal Beautiful." In the same way, "@Seeta is Noww Hardworkingg" becomes "Seet now hardworking."

2. Feature Extraction
   After pre-processing, the enhanced dataset has a number of distinct properties. The feature extraction method takes an attribute (adjective) from a dataset and extracts it. Later, this adjective is used to indicate positive and negative polarity in a sentence, which is useful for using the unigram model to determine people's opinions [15]. The adjective is extracted and separated using the Unigram model. It ignores the words that come before and after the adjective in the sentences. Only Beautiful is extracted from the sentence in the above example, i.e. “tajmahal Beautiful” using the unigram model.

3. Training and classification
   For solving classification problems, supervised learning is a useful technique. We used various supervised techniques to achieve the desired sentiment analysis result in this work as well. The three supervised techniques, nave bayes, maximum entropy, and support vector machine, were briefly discussed in the following paragraphs, accompanied by semantic analysis, which was used in conjunction with all three techniques to compute similarity.

- Naive Bayes
  Because of its versatility, it has been used in both the training and classification stages. It's a probabilistic classifier that can learn the pattern of evaluating a series of categorised documents. It compares the contents of the documents to a list of terms to classify them into the appropriate category [16].

$$P(A \mid B) = \frac{P(B \mid A)\ P(A)}{P(B)}$$

A and B are occurrences, and P(B) is a function. 0. Essentially, we're looking for the likelihood of event A if event B is real. Proof is often referred to as Event B. The priori of A is P(A) (the prior probability, i.e. Probability of event before evidence is seen). The proof is a value assigned to an undefined instance's attribute (here, it is event B). P(A|B) is the a posteriori likelihood of B, or the probability of an occurrence after seeing proof.

- Maximum entropy
  The entropy defined on the conditional probability distribution is maximized by maximum entropy. It also takes into account overlap and is similar to logistic regression in terms of determining distribution across groups. It also adheres to a set of function exception rules [17]. It uses the same processes as the nave bayes discussed earlier to determine the polarity of sentiments.

- Support vector machine
  The support vector machine analyses the data, defines the decision boundaries, and performs computations in input space using kernels. The input data consists of two sets of m-dimensional vectors. Then each piece of data expressed as a vector is assigned to a specific class. The goal now is to find a margin between two classes that is not connected to any text. The margin of the classifier is defined by the distance; maximizing the margin reduces indecisive decisions. SVM also supports classification and regression, which are useful in mathematical learning theory, and it aids in understanding the variables that must be considered in order to fully comprehend it [18].
- Semantic Analysis
  We used semantic analysis after the training and classification. The Word Net database, where each term is correlated with the others, is used for semantic analysis. This database is made up of interconnected English phrases. When two words are semantically identical, they are close to each other. We may

establish synonyms such as similarity in more detail. In the ontology, we map terms and investigate their relationships. The main task is to look through the stored documents for terms and compare them to the words that the user uses in their sentences. For users, it is thus beneficial to show the polarity of sentiment. For instance, in the sentence "I am happy," the adjective "happy" is selected and compared to the stored feature vector for synonyms. Assume two words: 'glad' and 'satisfied,' which are quite close to the word 'happy.' Following the semantic review, the word ‘glad' has replaced the word ‘happy,' resulting in an optimistic polarity.

IV IMPLEMENTATION AND RESULT

The naive bayes, maximum entropy, and help vector machine were trained and classified using Python and Natural Language Tool Kit. We used a total data set of 19340 bytes, with 18340 bytes used for training and 1000 bytes used for testing. Figure 2 is for preparation. Show the overall process movement.

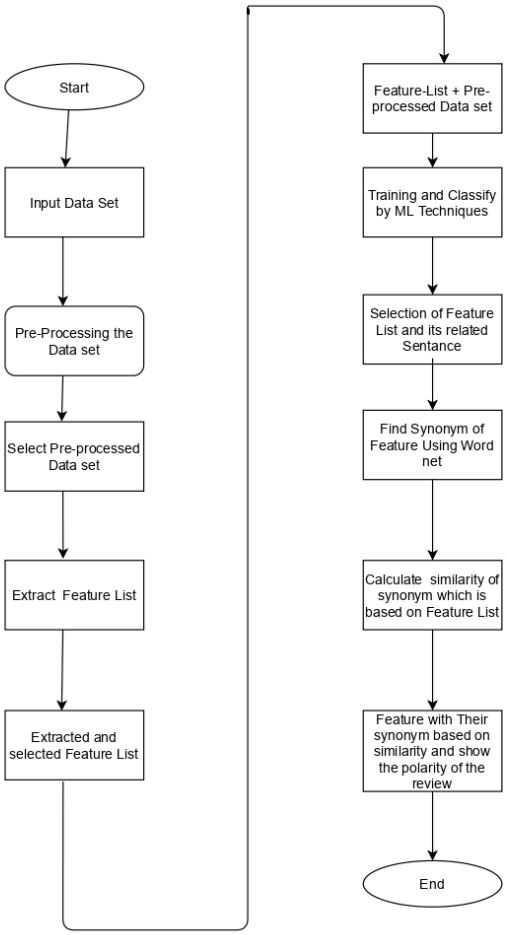

Fig. 2. Flow Diagram of the proposed methodology

The pseudocode definition of the method is shown. in figure 3.
Input: Labeled Dataset
Output: with synonyms and similarity between terms, positive and negative polarity
Step-1 Pre-Processing the tweets:
Pre-processing ()
Remove URL
Remove special symbols
Convert to lower
Step-2 Get the Feature Vector List:
For m in words:
Replace two or more words
Strip:
If (m in stopwords)
Continue
Else:
Append the file
Return feature vector
Step-3 Extract Features from Feature Vector List:

For word in feature list
Features=word in tweets_words
Return features
Step-4 Combine Pre-Processing Dataset and Feature Vector List
Pre-processed file=path name of the file
Stopwords=file path name
Feature Vector List=file path of feature vector list
Step-5 Training the step 4
Apply classifiers classes
Step-6 Find Synonym and Similarity of the Feature Vector
For every sentences in feature list
Extract feature vector in the tweets ()
For each Feature Vector: x
For each Feature Vector: y
Find the similarity(x, y)
If (similarity>threshold)
Match found
Feature Vector: x= Feature Vector: y
Classify (x, y)
Print: sentiment polarity with similar feature words

Fig. 3. Pseudo code of the process:

A. Results Analysis:

In this section, we compare the relative performances of nave bayes, maximum entropy, and support vector machine on three parameters: accuracy, precision, and recall

- Accuracy is expressed as a percentage and is calculated as follows:

$$\text{Accurcy} = \frac{Tp + Tn}{Tp + Tn + Fn + Fp}$$

- The recall ratio is calculated as: Recall positive (p) and Recall negative (n).

$$\text{Recall (P)} = \frac{Tp}{Tp + Fn}$$

$$\text{Recall (n)} = \frac{Tp}{Tp + Fn}$$

- Precision ratios are calculated as: Precision positive (p) and Precision negative (n).

$$\text{Precision (p)} = \frac{Tp}{Tp + Fp}$$

$$\text{Precision (n)} = \frac{Tn}{Fn + Tn}$$

In terms of precision and recall, Table 1, Table 2, and Table 3 display the output measurements of naive bayes, maximum entropy, and help vector machine dependent classifiers, respectively. Similarly, Table 4 displays the classifiers' accuracy results. Figure 4 depicts an overview of the accuracy of the three supervised learning methods as well as semantic analysis (WordNet). Figure 5 shows a comparison calculation based on the recall parameter. Figure 6 shows a similar comparison of measurements based on the precision parameter.

TABLE I. NAIVE BAYESIAN CLASSIFICATION MEASUREMENTS

Performance	Measures (%)
Positive Recall	91 %
Negative Recall	85 %
Positive Precision	49 %
Negative Precision	39 %

TABLE II. MAXIMUM ENTROPY MEASUREMENTS

Performance	Measures (%)
Positive Recall	86 %
Negative Recall	80.0 %
Positive Precision	40.4 %
Negative Precision	33.7 %

TABLE III. SUPPORT VECTOR MACHINE MEASUREMENTS

Performance	Measures (%)
Positive Recall	88.5 %
Negative Recall	83.6 %
Positive Precision	43.9 %
Negative Precision	35.8 %

TABLE IV. ACCURACY COMPARISON

Methods	Accuracy
Naive Bayes	88.3 %
Maximum Entropy	83.9 %
Support Vector Machine	85.5 %
Semantic Analysis (Word Net)	89.8

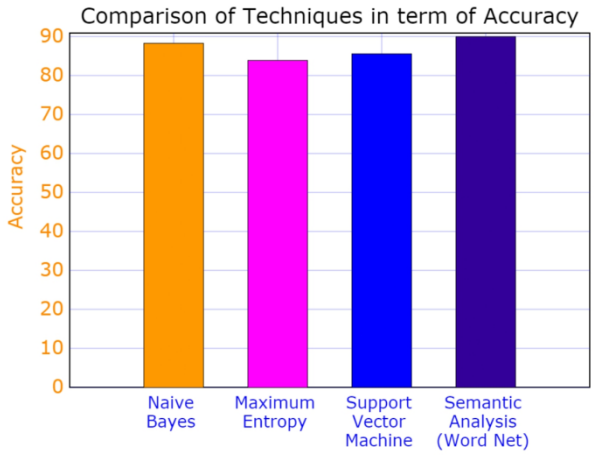

Fig. 4. Performance comparison of techniques in terms of accuracy

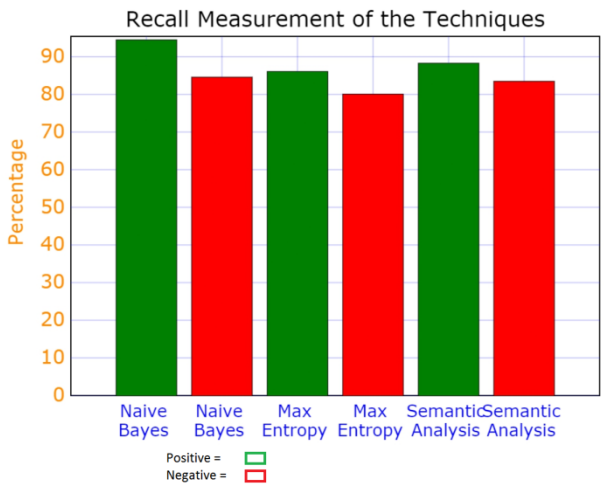

Fig.5. Measurements of positive and negative recall of the techniques

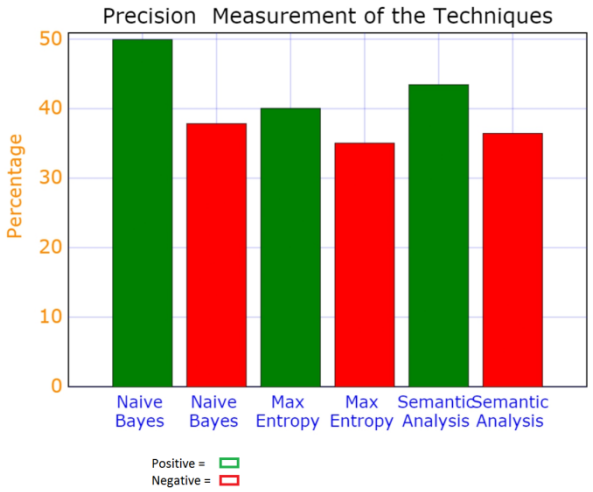

Fig.6. Measurements of positive and negative precision of the techniques

V. CONCLUSION

We proposed a collection of machine learning techniques with semantic analysis for classifying sentences and product reviews based on Twitter data in this paper. The main goal is to evaluate a large number of reviews using a pre-labeled Twitter dataset. The nave byes technique gives us a better result than maximum entropy, and SVM is subjected to a unigram model, which gives us a better result than just using it. When the semantic analysis WordNet is combined with the above technique, the accuracy increases to 89.9% from 88.2%.The training data set can be expanded to improve the feature vector-related sentence recognition process, as well as WordNet for analysis summarization. It will provide a better visual representation of the information, which will be beneficial to the users.

REFERENCES

[1] R. Feldman, " Techniques and Applications for Sentiment Analysis," Communications of the ACM, Vol. 56 No. 4, pp. 82- 89, 2013.

[2] Y. Singh, P. K. Bhatia, and O.P. Sangwan, "A Review of Studies on Machine Learning Techniques," International Journal of Computer Science and Security, Volume (1) : Issue (1), pp. 70-84, 2007.

[3] P.D. Turney," Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424, July 2002.

[4] Ch.L.Liu, W.H. Hsaio, C.H. Lee,and G.C.Lu, and E. Jou," Movie Rating and Review Summarization in Mobile Environment," IEEE Transactions on Systems, Man, and Cybernetics, Part C 42(3):pp.397-407, 2012.

[5] Y.Luo,W.Huang," Product Review Information Extraction Based on Adjective Opinion Words," Fourth International Joint Conference on Computational Sciences and Optimization (CSO), pp.1309 – 1313,2011.

[6] R.Liu,R.Xiong,and L.Song, ”A Sentiment Classification Method for Chinese Document,” Processed of the 5th International Conference on Computer Science and Education (ICCSE), pp. 918 – 922, 2010.

[7] A.khan,B.Baharudin, ”Sentiment Classification Using Sentence-level Semantic Orientation of Opinion Terms from Blogs,” Processed on National Postgraduate Conference (NPC), pp. 1 – 7, 2011.

[8] L.Ramachandran,E.F.Gehringer, ”Automated Assessment of Review Quality Using Latent Semantic Analysis,” ICALT, IEEE Computer Society, pp. 136-138, 2011.

[9] B.Agarwal,V.K.Sharma,andN.Mittal,”Sentiment Classification of Review Documents using Phrase Patterns,” International Conference on Advances in Computing, Communications and Informatics (ICACCI),pp. 1577-1580, . 2013.

[10] J.Zhu, H.Wang, M.Zhu, B.K.Tsou, and M.Ma,,” Aspect-Based Opinion Polling from Customer Reviews,” T. Affective Computing2(1):pp. 37- 49, 2011.

[11] M.Karamibekr,A.A.Ghorbani,”Verb Oriented Sentiment Classification,” Processed of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol (1): pp. 327-331, 2012.

[12] A. Neviarouskaya, H.Prendinger, and M.Ishizuka,” SentiFul: A Lexicon for Sentiment Analysis,” T. Affective Computing 2(1), pp.22-36, 2011.

[13] L.Liu, X.Nie,and H.Wang,” Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis,” Processed of the 5th Image International Congress on Signal Processing (CISP), pp. 1620 – 1624, 2012.

[14] R. Srivastava, M. P. S. Bhatia,” Quantifying Modified Opinion Strength: A Fuzzy Inference System for Sentiment Analysis,” International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1512-1519, 2013.

[15] C. Tillmann , and F. Xia, “A phrase-based unigram model for statistical machine translation,” Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL, pp.106-108, 2003.

[16] B.Ren,L.Cheng,” Research of Classification System based on Naïve Bayes and MetaClass,” Second International Conference on Information and Computing Science, ICIC ’09, Vol(3), pp. 154 – 156, 2009.

[17] C.I.Tsatsoulis,M.Hofmann,”Focusing on Maximum Entropy Classification of Lyrics by Tom Waits,” IEEE International on Advance Computing Conference (IACC), pp. 664 – 667, 2014. [18] M.A. Hearst,”Support vector machines,”IEEE Intelligent Systems, pp. 18-28, 1998.