Journal of Engineering Sciences

Vol 12, Issue 06, JUNE /2021
ISSN NO:0377-9254

SENTIMENT ANALYSIS OF TWEETS USING SUPPORT VECTOR MACHINE FOR ANALYZING WOMEN SAFETY

[1] *Yamuna pentapalli* [2]*Mr.L.Venkateswara Kiran,*[3]*Mr.K.Praveen kumar*

[1]***PG student,*** Dept of Computer Applications, Godavari Institute of Engineering and Technology(Autonomous),Rajahmundry , AP

[2] ***Assistant Professor,*** Dept of Computer Applications, Godavari Institute of Engineering and Technology(Autonomous),Rajahmundry , AP

[3]***Assistant professor,*** Dept of Computer Applications, Godavari Institute of Engineering and Technology(Autonomous), Rajahmundry , AP

Email:[1] yamunapentapalli1997@gmail.com,[2]lvkiran@giet.ac.in,[3]Praveenkumar@giet.ac.in

ABSTRACT: In India woman safety is foremost problem especially at night times, they are very frightened to come out from their houses. In order to give the protection to women, a systematic control is needed to avoid those scary events against the woman. With the references of sites of online media and their applications such as twitter, the required centre places around the web based media are observed to maintain security environment to the woman. Because of its unformatted data, spelling mistakes, slangs, limited size and abbreviations, it is more difficult to find the twitter data. Natural Language processing is used to find the sentiment analysis of unformatted twitter data. By using the support vector machine (SVM) produced algorithm of machine learning the classification of sentiments of the data from the tweets is processed. Several steps of pre-processing techniques are done by the researches which give the input to SVM machine learning classifier. For every sentiment average and classification the opinions are obtained by considering the F-measure and accuracy of classification. By using this proposed algorithm approach the security to women can be achieved.

KEYWORDS: Women safety, support vector machine (SVM), sentiment analysis, machine learning classifications and tweets.

I. INTRODUCTION

Throughout India women are facing the sexual harassment. The research has been conducted towards to this problem in some states and the result of this survey gives shocking facts that in every 4 in 1 women experience this forced sexual violence by the closed partner also those who believed. Only 30 percent of women can report this crucial act. In India 60% of women are in unprotected conditions while travelling in public places. Little girls are also facing this violence from their neighbours in some places. In public places like bus stands, temples and working places like offices women may feel safe and free. Women sexual violence can be reduced by sentiment analysis. Face book, twitter and instagram are the various social networks which are used by the public and their opinions are shared by these networks and these opinions are called as sentiments [1].

Most of the women use this social media to share their experience. The overall opinions of social media are extracted by the sentiment views central analysis and users exact opinions are is noticed [2]. This is the main task of sentiment analysis and it can extract the opinions as self-opinionated method [3]. In the business also this sentiment analysis is used widely. Especially from the user's point of view the selection of topic or product can be taken to divide the opinions present in the given data in different ways. In the form of online forums and blogs the data is expressed. So

researchers are kept attention towards to the social media because these forms are producing the maximum opinions to the products. By analyzing these opinions the polarities are obtained as negative, positive and neutral. The mixed review can be obtained from the several opinions of sentiment words which are very useful to users to express their feelings towards to the product or service [4].

Ninety-nine percent of world is surrounded by information only. Different products opinions are placed in the different micro blogging sites. If the correct means are not utilized then the individual user's opinions are wasted. To use this all opinions which are raised by the every user can be utilized in order to make huge productivity. Therefore this is a source of new technique or application [5].

II. MACHINE LEARNING IN SENTIMENT ANALYSIS

Large information is present in the social media or web which is valuable to the several industries or organizations so there is an attention in these data extraction form the huge data. So the method of study is raise to extract information is called sentiment analysis. This sentiment analysis is having the different names as data mining, opinion extraction etc. Even though small differences present in the operation processes between the names. Traditional techniques of survey can be highly influenced which were taken from the users individually before the automatic process of sentiments from the data. So from the individual data hundreds and thousands of hidden opinions in the reviews, user's posts and blogs a new automatic system is raised [6]. Movie reviews, politics, product reviews, recommender system etc are the several applications which use the sentiment analysis [7]. The organizations are changed in consequent with the different characteristics and opinions about the product. Some government schemes are changed accordingly on the bases of opinions about the particular party. Lexicon based and machine learning based are the two techniques which are widely used to find the sentiment analysis [8] [9].

In addition with Lexicon based and machine learning based there is another method to find the sentiment analysis that is hybrid approach [10]. Due to its metric capacity and language options Machine Learning approach (ML) is used widely to do sentiment analysis. Depending on the sentiment lexicon the process of the Lexicon-based Approach is done. By collecting of various sentiment words or terms then after the process is compiled. The sentiment polarity of the text can be mentioned after dividing the data into corpus by this approach. Primarily based approach and dictionary-based approach are uses the math methods and language punctuations. The lexicon method plays a key role in the Hybrid Approach which combines the common sentiments to the lexicons [11].

Day to day the communication is maintaining a key role in life. In different applications such as science, education, business Machine learning is one major technology which is used widely. With the previous data the computer is trained and which gives the calculations to the inputs accurately in the machine learning techniques. The instructions are given to the computer by this machine learning approach which can allow the data to learn to avoid the programmer to give the instructions step by step process. So manually cannot done complicated works can be done by this process. Unsupervised, Supervised and

Reinforcement are three different types of algorithms involved in the machine learning [12]. The dependent variables are predicted by the supervised algorithm from the finite set of predictors and independent variables. Support Vector Machine (SVM), KNN, Linear regression, Logistic Regression, Naïve Bayes and etc are the different classifications in the supervised learning algorithm.

2.1 Naïve Bayesian

By using Bayesian theorem this Naïve bayes classification performs the classification. So it is powerful classification by comparing with other. Depending on the arrived probabilities the beliefs are variables which are independent to one with others. By taking the variance and mean of variables in account this Naïve bayes classification is takes place. It can handle the large data sets having the complicated parameters also. It can build the classification structure. One parent and many children with acyclic graphs are used to design the Naïve bayes algorithm [13].

2.2 Logistic Regression

In numerical algorithmic classification methods this Logistic Regression is most effective method. The dataset can be analysed with the generating binary output by using the statistical method. So it can easily implement the binary classification methods. By identifying the relation between independent features and one dependent variable and can expects the binary variable by using this Logistic Regression. By taking the output variables into consideration this Logistic Regression are three types as ordinal, multinomial and binomial [14].

2.3 K-Nearest Neighbour

It is a simple algorithm while comparing with other algorithms. In this category wise the operation is processed. At first step the new data is compared with the already existing data sets, and then it assigns the new data to the related existing datasets. It is non parametric algorithm so this cannot take any guess on the data. The new data can be categorised by using this K-NN algorithm easily [15].

III. PROPOSED ARCHITECTURE:

Framework of Women safety uses the SVM based sentiment analysis is shown below figure (1). This operation is done in three steps. In the first step the impurities are filtered and correction is needed in this step. To make normalize input tweets in this step stemming, correction of spellings, stop words removal are defined and including these operations tag tokens, negative aspects and positive aspects are defined and negation also maintained in this. In the next step the statistical features are generated by the filtered text from the first step. The transforming input is tested and trained according to the equivalent feature set. In the final step the sentiment output is obtained by the features which are processed by the support vector machine.

For the individual stage, the predictive decision is assigned to select the SVM machine. These tweets are collected from the web. For twitter data extraction, hashtags have to be provided, so that data relevant to those hashtags will be retrieved. The hashtags uses as: Believe Women, Why I Didn't Report, Women's Reality, MeAt14, I'm with Her, Me Too, To the Girls, Why I Stayed, sexual harassment, sexual assault, rape, molestation.

3.1 Word Tokenization: The text is converted into the form of tokens before converted into vectors is called tokenization. The insignificant tokens are eliminated by using this process. Let the example of a document which converts the sentences into words then in this case reviews of words are tokenized.

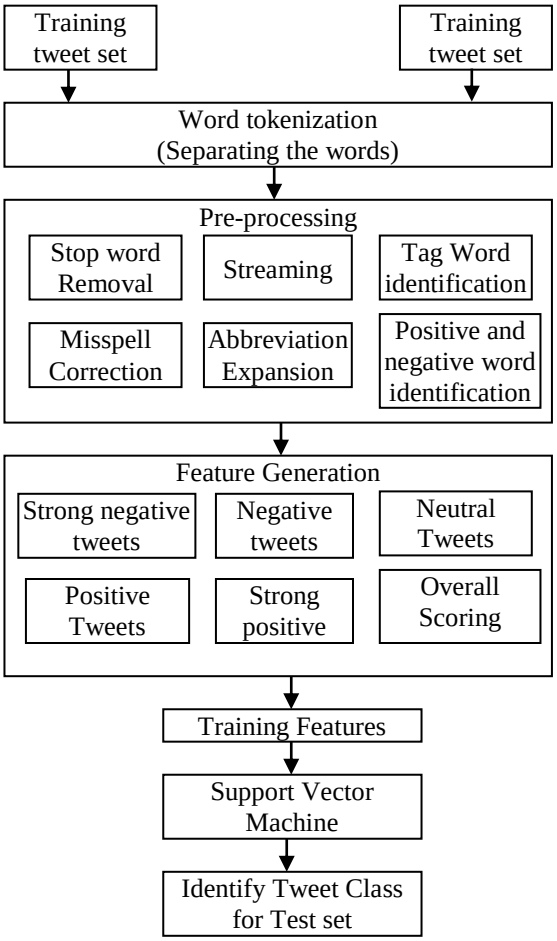

FIG. 1: FRAMEWORK OF SVM BASED SENTIMENT ANALYSIS

3.1 System architecture

3.1 Pre-processing: In this step several operations are performed such as removing of stop word, expansion of abbreviation, listing the positive words in the tweet, listing the negative words in the tweet and negation handling. For the streaming Porters algorithm is used. Stop words are placed in the sentences. Those are having any meaning towards to the paragraph or data and it cannot have the phrase meaning also in which is not having any sentiment. Stemming is a process of removing inflectional words which is affixes, for example playing-play, studies- study. Stemming works on some particular language mainly English and Spanish.

The data base is created for the misspell correction, abbreviation expansion and stop words. After the tokenization the tweets with filtered list is obtained. This filtered list contained the @ tags and removed with filtered tweets. In every tweet the tag count as feature generation this @ tags are used. In the tweet the negative adjectives are in the Negative list likewise positive adjectives are in the positive list.

3.2 Feature Gathering: The positive opinions or negative opinions are clearly given by the terms, words or phrases in the given data of opinion data mining for the features. Therefore the position of words in particular sentence having the higher priority by comparing with the same words in other sentences. Feature selection uses the different methods. Some as syntactic are based on the adjectives like position of syntactic. Some as univariate based on specific category of feature relations. Some are multivariate which uses the decision trees and genetic algorithms based on the subsets of features. The importance of every feature is processed in different ways in order to add the specific weight to the data. Feature Frequency (FF), Feature presence (FP) and Term Frequency Inverse Document Frequency (TF-IDF). In the document number of occurrences is present in the FF

and feature present or absent representing as 1 or 0 can be taken by the FP. In this work, the features are extracted from the documents by using the bag of words. In order to train the algorithms of machine learning these extracted features are used. It makes a jargon of the apparent multitude of novel words happens in all the reports in the preparation set. Bag of words features containing term frequencies of each word in each document, i.e. the number of occurrence and not sequence or order of words matters. For the feature generation adjectives list are used. The characteristics of the all adjectives give the list of negative score, positive score and overall rating. The tweets are filtered for the classifiers then the different features are produced. For learning the classifiers used the several features as:

Word count:-After filtration the number of words is counted.

Tag count:-each tweet using @ tags.

Negative word count:-having of each tweet containing negative words.

Positive word count:-having of each tweet containing positive words.

Positive score:-it is obtained by adding all positive adjectives score.

Negative score:-. it is obtained by adding all negative adjectives score.

Score:-by adding Positive score-Negative score for each tweet.

Message class 0: for negative tweets
1: for neutral tweets
2: for positive tweets

3.3 Support Vector Machine (SVM):

SVM is abbreviated as supervised machine learning. This algorithm is widely used in problems classifications. The separating hyper plane is defines the Support Vector Machine classifier. This hyper plane makes good differences in the two classes and by the discovery of hyper plane the implementation of classification is done. It performs very well with a limited amount of data. The extracting features after the data pre-processing which is called as bag-of-words. Sentiment classification values are can be calculated by using the SVM. For getting the feature, the words are embedded and generation of feature vector is done which gives the data to classification process. By finding the hyper plane the classification process is done in the SVM and the classes are sketches in the n-dimensional space are separated by the hyper plane. The analysis from the SVM is predicted weather it is negative or positive.

IV. RESULTS

Different characteristics are available in this data such as people frequency, destination, time, source time, time to travel, police station availability, tier, residence level and presence of bars. For example, if a woman can starts her journey which takes more time from source to destination as 6 to 7 hours and timings that means night or day times gives the idea about that journey is it safe or not can be predicted. At night times the journey of alone is risky which may cause some terrific actions on her. So in those situations the described paper can predict the travelling place is safe or not according to the variables of police station availability, people frequency and bar presence.

Yes or no values are given to the bar availability and police station presence. That destination place is inner, outer or middle of the city can be predicted by the tier. The values of evening, morning, nights and afternoons can be analysed by the time attribute. The values are given as low, high and medium for indicating the frequency of people. Safe or unsafe are given by the targeted attribute class. For easy visualization all the data is converted into numerical values. 75 percent of dividing

data is trained first and the remaining data is tested. The existing algorithms are used by the training data to identify the pattern. By comparing the actual outcomes with predicted outcomes the model can be evaluated by using the testing data. Pre-process the data is needed before the training and testing. The missing data can be removed in this pre-processing stage. SVM works better in numerical data so all the data can be converted into the numerical values. The following parameters are used to measure the performance of proposed frame work:

• Accuracy:

Accuracy can be defined as the ratio between the true values (both positives and negatives) to the analysed total number of documents by classifiers.

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FP + FN}$$

• Precision:

Precision is defined as the ratio of classified positives in sentiments to the number of documents having positive sentiments in the given text corpus.

$$Precision\ (P) = \frac{TP}{TP + FP}$$

• Recall:

Recall is defined as the ratio of number of documents positive classified to the actual number of documents having positive sentiments in the given text corpus.

$$Recall\ (R) = \frac{TP}{TP + FN}$$

• F1 Score:

Precision and recall harmonic mean can be defined by the F1score.

$$F1\ Score = 2 \times \frac{P \times R}{P + R}$$

Where, TP, FP are abbreviated as true positives, false positives and TN, FN are abbreviated as true negatives, false negatives. The results are obtained in terms of TP, FP, TN and FN. The below shows the confusion matrix as,

Table 1: CONFUSION MATRIX OF SENTIMENTS

Class	Negative	Neutral	Positive
Negative	82	16	20
Neutral	12	49	15
Positive	18	18	80

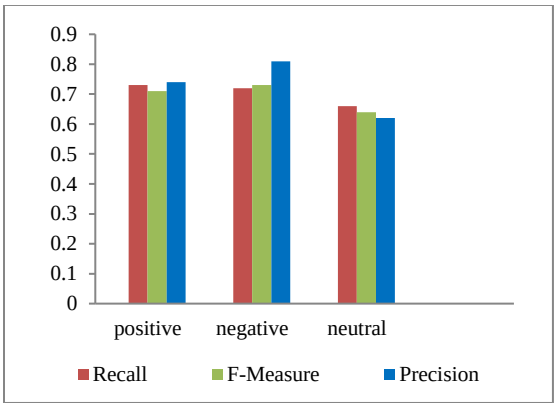

Fig. 4: PERFORMANCE COMPARISON OF DIFFERENT SENTIMENTS

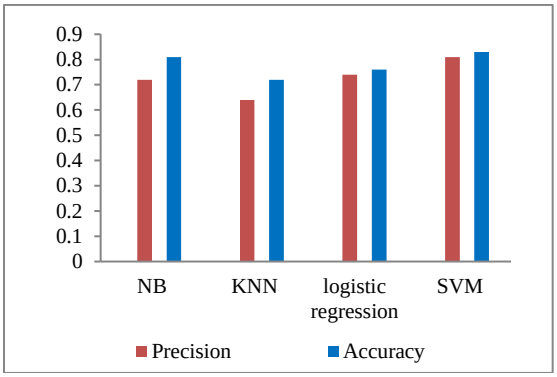

Fig. 4.1: ACCURACY ANND PRECISION COMPARISION OF CLASSIFIERS

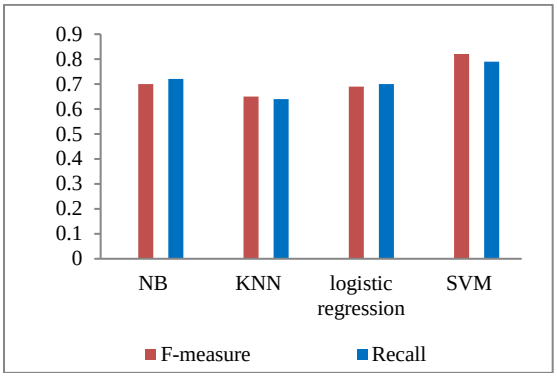

Fig. 4.2: RECALL AND F-MEASURE

COMPARISION OF CLASSIFIERS

The comparison values of positive, negative and neutral classes with the help of Recall, F-measure and precision can be calculated.

V. CONCLUSION

The goal of this paper was to extract tweets related to women harassment in Indian states and find out which are not safe for women. By using several stages of performance the proposed frame work can be done. Those stages are pre-processing stage, features generation stage and classifiers learning stage. So the results are obtained in terms of positives, negatives and neutral sentiments. If the neutral tweets are significantly top, it means that people have a lower interest in the topic and are not willing to have a positive/negative side on it. By using the accuracy and f-measure the analytical values are calculated. The comparative observations are taken against the SVM, KNN, NB and logistic regression methods. On comparing the accuracies of Random Forest, Support Vector Machine (SVM) and Naïve Bayes algorithms it has been found that the Support Vector Machine classifier gives out the improved accuracy and f-measure of tweet class prediction.

VI. REFERENCES

[1] Zhitao Wang, Chengyao Chen, Wenjie Li, "Tracking Dynamics of Opinion Behaviors with a Content-Based Sequential Opinion Influence Model", IEEE Trans. on Affective Computing, Vol: 11, Issue: 4, 2020

[2] Joarder Kamruzzaman, Rajkumar Das, Gour Karmakar, "Opinion Formation in Online Social Networks: Exploiting Predisposition, Interaction, and Credibility", IEEE Trans. on Computational Social Systems, Vol: 6, Issue: 3, 2019

[3] Guijin Tang, Xi Shao, Bing-Kun Bao, "Personalized Travel Recommendation Based on Sentiment-Aware Multimodal Topic Model", IEEE Access, Volume: 7, 2019

[4] Francisco Jurado, Ruth Cobos, Alberto Blázquez-Herranz, "A Content Analysis System That Supports Sentiment Analysis for Subjectivity and Polarity Detection in Online Courses", IEEE Revista Iberoamericana deTecnologias del Aprendizaje, Volume: 14, Issue: 4, 2019.

[5] Sher Muhammad Daudpota,Irum Sindhu, Mohammad Nurunnabi, Kamal Badar, Maheen Bakhtyar, Junaid Baber, "Aspect-Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation", IEEE Access, Volume: 7, 2019

[6] Yanni Liu, Liping Yu, Dongsheng Liu, Liming Wang, "Research on Intelligence Computing Models of Fine-Grained Opinion Mining in Online Reviews", IEEE Access, 2019.

[7] Wei ssZhao, Long Chen, Deng Cai, Xiaofei He, Quan Wang, Beidou Wang, "Weakly-Supervised Deep Embedding for Product Review Sentiment analysis", IEEE Trans. on Knowledge and Data Engg., Volume: 30, Issue: 1, 2018

[8] Donglin Cao, Fuhai Chen,Yue Gao, Rongrong Ji, Jinsong Su, "Predicting Microblog Sentiments via Weakly Supervised Multimodal Deep Learning", IEEE Trans. on Multimedia, Vol: 20, Issue: 4, 2018

[9]Veenu Mangat, Harpreet Kaur, Nidhi, "A survey of sentiment analysis techniques", 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2017
[10]Marco Guerinil,Lorenzo Gatti, Marco Turchi, "SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis", IEEE Trans. on Affective Computing, Volume: 7, Issue: 4, 2016
[11]Mehrdad Jalali, Shokoufeh Salem Minab, Mohammad Hossein Moattar, "A new sentiment classification method based on hybrid classification in Twitter", 2015 International Congress on Tech., Communication and Know. (ICTCK), 2015
[12]Wenyan Gan, Hailong Zhang, Bo Jiang, "Machine Learning and Lexicon Based Methods for Sentiment Classification: A Survey", 2014 11th Web Information System and Apps Conference, 2014
[13]Sherly Novianti Thahir, Liza Wikarsa, "A text mining application of emotion classifications of Twitter's users using Naïve Bayes method", 2015 1st International Conference on Wireless and Telematics (ICWT), 2014
[14]Hao Yu, Yu Mao, Xiaojie Wang,Muyuan Xi,"Semi-supervised logistic regression via manifold regularization", 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, 2011
[15]Yi-Bing Geng,Hai-Bing Ma, Jun-Rui Qiu, "Analysis Of Three Methods For Web-Based Opinion Mining", 2011 International Conference on Machine Learning and Cybernetics, Vol: 2,