Sentiment Analysis of Tweets using Machine Learning Approach

Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma, Sarthak Mendiratta
Jaypee Institute of Information Technology
JIIT Sec-62
Noida, India
megha.rathi@jiit.ac.in

Abstract−Microblogging 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. The main emphasis of this research is on the classification of emotions of tweets' data gathered from Twitter. In the past, researchers were using existing machine learning techniques for sentiment analysis but the results showed that existing machine learning techniques were not providing better results of sentiment classification. 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 merging Support Vector Machine with Decision Tree 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.

Keywords— sentimental analysis; social media; Twitter; Hybrid; Decision Tree; Adaboosted Decision Tree; SVM.

I. INTRODUCTION

Currently, social networking sites are at an all-time high, so from there, a large amount of data is generated. IOT has molded the way social networking sites users express their feelings and opinions. Twitter is a reservoir for a large amount of data. So, this data is extremely useful for predicting results of political activities, new initiatives led by government, or research and deciding on what content to share with the audience. Input to our model is the raw data extracted from tweets. For the same, we automate the process of tweet extraction and categorizing it into two categories i.e. positive or negative. The content in twitter generated by the user is about different kinds of products, event, people and political affairs [2].

Performing sentiment analysis on tweets is considered best due to the following reasons:

1. Tweets are abstract in nature and its length is 280 characters.
2. Analysis in real time can be done.
3. A vast variety of tweets for performing the analysis.

II. ALGORITHMS STUDIED

Firstly, we will be classifying the tweets on the basis of Support Vector Machine (SVM) and adaboosted Decision Tree individually. The tweets i.e. string input will be taken which will be converted to numeric type by TF-IDF. Then a hybrid technique will be applied by feeding the outputs obtained earlier as the input to the Decision Tree.

A. Tf–idf

In context of information retrieval, tf–idf or TFIDF (term frequency–inverse document frequency) is used which provides a statistic of how a particular word is crucial for the given document dataset. Tf-idf uses weights for text mining and information retrieval and its value is directly dependent on the number of a particular word within the given dataset and we count the frequency of each word in the given document for a better information retrieval that regulates more frequently appearing words generally. Tf-idf weight is used in the search query when searching for a given document relevant to the given user query and is used by the search engines in the ranking of the document. Term Frequency (tf) describes how frequently a word appears in a document. Inverse document frequency (idf) describes how important a particular word is with respect to the given document. Consider a dataset D, word w and document records dϵ D, we tend to compute:

$$w_d = f(w,d) * \log(|D|/f(w,D)$$

Where D is the collection of records, f (w, d) is the frequency of how many times a term appears in the given document, w_d represents the importance of a given term. One of the best classification methods is worked out by adding the tf–idf or every search query word. Besides this, other subtle classification methods are variations of this easy model. [3]

B. Support Vector Machine(SVM)

Support Vector Machines [4] employ the technique from computational learning theory which basically aims at

978-1-5386-6835-1/18/$31.00 ©2018 IEEE

reducing the structural risk. Finding the decision boundary that maximizes the distance between two classes is the basic principle of SVM. The vectors that define this decision boundary are termed as support vectors [5]. SVM algorithm builds the classification model that assign test examples to one of the predefined class categories making it a non-probabilistic linear classifier. The basic principles of SVM involves three steps (1) finding out the optimal decision boundary which maximize the distance between two classes (2) enhancing the same (1) for non-linear separable problems (3) map data to high dimensional space so that data is classified easily for linear separable and non-linear separable cases [6].

C. Decision Tree

A decision tree is basically used to represent choices and their subsequent results in the form of graphs. Nodes of the graph represent an event and edges represent decision condition. The complex problems of branches are solved by segmenting it. A decision tree model is generated with training data and some sets of validations are used to check and improve the performance of decision tree model. It is the work of the decision tree to clarify all the branches which give us the answer to a complex solution. The most popular classification technique is the DECISION TREE in data mining [7]. In a tree form they form some of the classification rules, and have several other advantages when compared to other techniques.

- Easily understood because of it's the simplicity of its presentation
- Decision tree can be applied to any data types like nominal, ordinal, numerical, etc is punctuated within the parentheses.)
- Test data classify very fast with the help of decision tree algorithm.

  C4.5 tree developed by Ross Quinlan is one of the most commonly used and oldest techniques of decision tree algorithms [6]. The main idea here is to create a tree with training data along within cooperating information entropy concept [8].

D. ADABOOSTED DECISION TREE

Adaboost is a machine learning algorithm, it is a solution to the problem created by "boosting" weak classifiers into strong classifiers, which can be done by weighting their respective outputs. The input to the boosting algorithm is the training data in the form of t (x1, y1), (xm, ym) where xi represents instant space X and yi represents class label set Y. Let say Y = {0, 1}. AdaBoost algorithm recursively iterates the basic learning technique with t iterations denoted as 1 to t. Distributing weight set over the entire training dataset is the main objective of boosting algorithm. In the initial phase, if Dt (i) is the distributed weight of training dataset I on t iterations, all weights are equally distributed and, in every iteration, it gradually increases the weight of incorrect classified tuples and the weak learners are focused more in the training examples.

III. ACCURACY

Classification accuracy of the predictive model is defined as the count of test cases out of every 100 test cases which were classified correctly by the model. The accuracy of Level 1 Algorithm 1 which is adaboosted decision tree is 67, Accuracy of Level 1 Algorithm 2 which is SVM is 82%, and accuracy of the third algorithm which is hybrid of a decision tree, adaboosted decision tree and SVM (Level 2) is 84%.
Analysis results show Positives (1), False (0).

IV. PROPOSED METHODOLOGY

Proposed hybrid model for sentiment classification is basically a three-step process:
In the first step of the proposed model, three datasets have been incorporated for training the classifier.
We have implemented various machine learning algorithms for classification of tweets. The datasets being used and their sources are explained below [14]:

1) Stanford Sentiment140
   Size: 1,600,000 tweets
2) Polarity Dataset
   Size: A collection of processed movie reviews (1000 positive and 1000 negative reviews)
3) University of Michigan
   Size: 7086 sentences from social media (not necessarily twitter).

Which is followed by the PREPROCESSING of data: -
i. Lower Case - Tweets are converted to lower case
ii. URLs - Convert www.* or https? ://* to 'URL'
iii. @username - Convert username to '__HANDLE'
iv. #hashtag - Hashtags mostly contain useful information, so we are replacing hashtag with its word written without the hash. E.g. #Apple is converted to 'Apple'[15]
v. Trimming the tweet
vi. Repeating words: In informal conversations, a person often uses repeating characters, such as "I'm happyyyyy". We are replacing characters with a frequency greater than two with that character repeated twice so that the result for above would be "I'm happyy".
vii. Emoticons: We identify a set of emoticons and replace them with the representative sentiment i.e. 'positive' or 'negative'. E.g. ':)' is replaced by 'positive'. In this step, we define a process for segregating tag tokens, positive aspects and negative aspects from messages. [13]
viii. Stemming: Stemming algorithms are used to search the "root" or base word of a given word. We have used the Porter Stemmer. [16]
viii. Tuning of Parameters: Tuning of parameters was done for improving the execution of the SVM classifier. The following parameters are found to give the most effective results on the cross-validation set (20% of the Training Corpus) without compromising much on the speed. [13]

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level

head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named "Heading 1," "Heading 2," "Heading 3," and "Heading 4" are prescribed.

ix. TfidfVectorizer:
The attributes (used under TfidfVectorizer) with their values are as below:
min_df was set to 5,
max_df was set to 0.95,
sublinear_tf was assigned True,
use_idf was assigned True,
ngram_range lies between (1, 2) [10]
In the second step, the preprocessed text is given as the input to the SVM and ADABOOST DECISION TREE classifier separately. Then, their corresponding outputs are saved. Then STACKNET HYBRID technique (stack net is a computational scalable and analytical meta-modeling framework) is used for the creation of this hybrid algorithm.
In the third stage, the outputs of SVM, ADABOOSTED DECISION TREE and the originally known output are served as the input to the DECISION TREE which gives the final output along with the accuracy.

Sequence no	**F-measure and Accuracy**		
	ALGORITHM USED	***F-measure***	***Accuracy***
1	SVM	82%	82%
2	Adaboosted D-Tree	67%	67%
3	Decision Tree	84%	84%

Table1: Value of F-measure and Accuracy

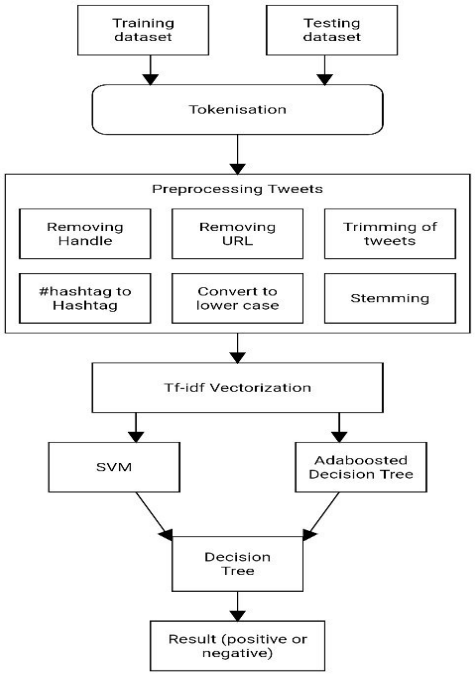

Fig 2: Complete flow diagram of the proposed system

V. EXPERIMENTAL RESULTS AND EVALUATION

In the proposed research work, we created a hybrid classification model by combining the Support Vector Machine (SVM), ADABOOSTED DECISION TREE and DECISION TREE to process and evaluate the twitter attributes and to classify the sentiments.

VI. CONCLUSION

In this paper, SVM, ADABOOSTED DECISION TREE and DECISION TREE based hybrid 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 and f-measure are used. The comparative observations are taken against the SVM, ADABOOSTED DECISION TREE and DECISION TREE.
The comparative results prove that hybrid model improved the overall classification accuracy and f-measure of sentiment prediction as compared to traditional existing techniques for classification.

REFERENCES

[1] V.Lakshmi, K.Harika, H.Bavishya, Ch.Sri Harsha, "SENTIMENT ANALYSIS OF TWITTER DATA," vol.04, February 2017. Link- "https://www.irjet.net/archives/V4/i3/IRJET-V4I3581.pdf"

[2] https://en.wikipedia.org/wiki/Twitter

[3] Using TF-IDF to Determine Word Relevance in Document Queries by Juan Ramos(Department of Computer Science, Rutgers University)

[4] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". Machine Learning. 20 (3): 273–297.

[5] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.

[6] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993

[7] Predicting Students Final GPA Using Decision Trees: A Case Study by Mashael A. Al-Barrak and Muna Al-Razgan

[8] https://www.lexalytics.com/lexablog/2014/sentiment-analysis-added-to-oxford-dictionaries

[9] A Short Introduction to Boosting by Yoav Freund and Robert E. Schapire (AT&T Labs, Research, Shannon Laboratory)

[10] A Hybrid Approach for SupervisedTwitter Sentiment Classification by K. Revathyand Dr. B. Sathiyabhama

[11] Using Objective Words in SentiWordNet to Improve Word-of-Mouth sentiment classification by Chihli hung and Hao-kai ling

[12] R. Parikh and M. Movassate, "Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques"

[13] A. Mudinas, D. Zhang, and M. Levene, "Combining lexicon and learning based approaches for concept-level sentiment analysis," Proc. First Int. Work. Issues Sentim. Discov. Opin. Min. - WISDOM '12, pp. 1–8, 2012.

[14] Pooja Kumari, Shikha Singh, Devika More, Dakshata Talpade, Manjiri Pathak, on"Sentiment Analysis of Tweets" dated April 2015

[15] Akshi Kumar and Teeja Mary Sebastian, on "Sentiment Analysis on Twitter", dated July 2012

[16] Michael Gamon. 2004. Sentiment classification on customer feedback.