Vol-07 Issue 01, January-2023 ISSN: 2456-9348
Impact Factor: 5.004

iJETRM
International Journal of Engineering Technology Research & Management

SENTIMENT ANALYSIS USING MACHINE LEARNING METHODS AND SEMANTIC RESEARCH TECHNIQUES ON TWITTER DATA

Nazir Ahmad ZAHIRI,
Shafqat Ur REHMAN
Ankara Yıldırım Beyazıt Üniversitesi
Computer Engineering Department

Abstract:
Analyzing the sentiment from the tweets of a user can give over-all visions into how the person is thinking. With the development of technology, the fast progression of social media on Web has improved day by day. A lot of People use the internet to precise their opinions and feelings in the form of reviews, tweets, blogs, comments, and postings them in a social network site. People and especially the governments are gradually using such content in these media for making decisions. Where Twitter is a platform which is most extensively used by people to direct their visions and sentiment in a tweet's method. Where a Sentiment analysis is the procedure of discovery whether a part of script is positive, negative or neutral. The goal this work is to implement sentiment analysis using machine learning methods to predict the sentiment and finally analyze the result in terms of precision, recall and f-score.

Keywords:
Sentiment analysis, social media, Machine learning, Twitter, Sentiment, Deep Learning, Neural Network, Polarity, SVM, Decision tree, Random Forest,

I. INTRODUCTION

Sentimental analysis (also known as opinion mining or emotion AI) is the use of natural language processing (NLP), computational linguistics, text analysis, and biometrics to systematically identify, citation, quantify, extract, and study sentimental states and particular or subjective info. Sentiment analysis is broadly applied to voice of the customer resources such as reviews and survey responses, online and social media, and also healthcare materials for applications that array from advertising to customer service to clinical medicine. In the last few years, Micro blogging websites have been advanced progressively and became general tool for communication between web users in recent years. Massive amount of data is produced through computer and mobile devices and most of the data is in written format [1]. Social media has been extensively used and become popular for sharing the information across the world or universe, social media acts as an energetic role for someone who wants to put their sentiments about any present or real actions. Social media that is also an important communication tool hence the age of Internet. It is proposed to help as an application to understand the opinions, thoughts and feelings expressed in the context of an online resource. These days numerous people have the habit of social networking sites to network with others and stay up-to-date with news and existing affairs. Such sites like (Twitter, Instagram, Facebook, etc.) deliver an opportunity for publics to express their views. For example, someone will vitally or automatically post their examinations or review online as soon as they see a movie that is where then it starts a series of discussions about the acting abilities shown in a movie. Such kind of data is the criterion for people to analyze, control the quality not only of any film, but also of other kind products, and to know whether its positive comment or negative comment. Such data or information become significant source of the internet data information. All this information can be valuable for not just e-commerce examination but that is also to analyze which sort of data or information that become popular in the society, one of them are Social Problem. Social problem is one of the evidences that is typically share in the social media. These data are need for the administration or government such as health ministry and Social or it is also essential for the non-government organization for social problem so that to prepare for the solution and avoiding the social problem. Twitter is seemed to be one of the main social websites [3]. To understand the information and handle the big data the procedure or algorithm and program is needed to process the information and comment data, imprisonment and

analyze the opinion of the social media users. On the other hand, sentiment analysis also effects users to sort whether the information about the product is suitable or not before they get it. Marketers and firms use this analysis to understand about their own products or services in such a way that it can be offered as per the user's demands.

All this information of tweets is usually loud, representative for shifting views, multi-topic information most importantly they are unstructured, and not clean and also in unfiltered format. Analyzing sentiment expressed still is not an easy mission. There are numerous difficulties in terms of tone, division/ polarity, terminology and tweet grammar. They seem to be highly informal and pseudo-grammatical. To analyze such kind of data there are two types of machine learning methods which are generally used for sentiment analysis, one is Supervised learning and the other is Unsupervised learning. Supervised learning are algorithms that requires exterior support. Which implies the input dataset is divided into train and test dataset. The train dataset has output flexible variable which demands to be predicted or classified. Simply we can say supervised learning is based on labelled dataset and thus the labels are provided to the model during the process. These labelled datasets are trained to produce reasonable outputs when encountered during decision- making. Where the unsupervised learning algorithms learns only some features from the data. Once new data is launched, it operates the earlier learned features to understand the class of the data. Simply Unsupervised learning does not consist of a category and they do not provide with the correct targets at all and therefore conduct clustering.

To aid us to better understand the sentiment analysis in a more improved way, in this paper we are going to be using different supervised machine learning methods. Among them are SVM (Support Vector Machine), Naive Bayes, k-Nearest Neighbors, XGBoost and different other algorithms to better achieve the results and understand the topic. Apart from these we are going to work on sentiment lexicon algorithm also where this algorithm compares each word in a tweet to a database of words that are labeled as having positive or negative sentiment.

In this paper, I will discuss the behavior of a sentiment analysis on Twitter datasets to see whether Twitter might help to advance choice making over analyzing people's approaches and feelings. The given data used for the following study are taken from Twitter real-time tweets for a specific time. Since the collected tweets are real data, it is unbalanced and unstructured. So the data are converted to an applicable format via preprocessing and NLP approaches. After the preprocessing method, I am going to apply machine learning and deep learning methods [11].

The contents of the paper are consisting as follow: Section I is the introductions of the research, the Section II introduce the related work or research on this field, the Section III introduce the problem and identification Section IV introduce the approach and methodology, the Section V introduce Implementation, and the Section VI introduce result and the conclusion.

II. LITERATURE REVIEW:

In last few years lot of work has been done in the field of "Sentiment analysis "by number of researchers. Where most of the research works have been discussed on sentiment analysis on social media's data. Meanwhile several techniques for sentiment analysis have been built and implanted over the years, none of them have been broadly tested with deep learning methods on the Twitter data. In this part of the section, I will discuss some of the previous research that has been done on these topics. The function of sentiment analysis is one of the eldest and most important processes for the development of NLP applications. Where different methods, such as SVMs and Naive Bayes, have been tried to solve such tricky problems. Today actions show that the sentiment analysis has reached success where it could pass through not just only positive and negative sentiment but also deal with the performance and sentiments for different topics and languages. In the research of sentiment analysis, researcher use numerous methods, techniques for predicting the social estimation and emotion trough the text and languages such as:

In the paper by S. M. Mohammad, S. Kiritchenko, and X. Zhu (2013) Applied SVMs for emotion analysis, composed optimistic, negative and neutral tweets from a number of sources, including the Sentiment140 database. The public extracted features from each message, like characters ngrams, emoticons, etc., for their classification and obtained a result of 69.02%. [2]

In the paper by Brett Duncan and Yanqing Zhang (2017) applied the neural network to categorize the sentiment on tweets. The average accuracy (number of correctly classified tweets divided by the number of incorrectly classified tweets) was 75.15 %. [4]

In the paper by P.D. Turney," Thumbs Up or Thumbs Down (2002) Applied unsupervised learning algorithm which classifies it as thumbs up or thumbs down review. Where they predict review by the average semantic orientation of

a phrase that contains adjective and adverb thus calculating whether the phrase is positive or negative and the result of 66.09%.[5]

In the paper by Zhang, X. and Zheng, X., 2016, July. Chinese text is measured for sentiment analysis which could be thought as two-classification problems to find the polarity of text, with positive and negative emotional leanings. They done data cleaning, word segmentation, removing stop words, feature selection and classification. The weights of features were calculated by the TF-IDF. For this work, SVM model is used for text representation and used SVM and ELM to analyzing text emotions. [6]

In the paper by R. Joshi and R. Tekchandani' (2016) applied the SVM (Support Vector Machine), Naïve Bayesian and the maximum entropy to compare the methods for twitter data analysis, for movie review from twitter for their dataset. [7]

In the paper by R. Liu, R. Xiong, and L. Song, (2010) has used a rule-based technique, built on Baseline and SVM for sentiment analysis of Chinese document level, which extract the overall document polarity of specific words by a sentiment word dictionary, and adjust it according to the context information. [8]

In the paper by J. Barnes, R. Klinger, and S. S. i. Walde (2017) approved parallel studies in the calculation of document classification methods. The group trained some models in the SemEval dataset for the Tweets classification and gained the strongest results using the Bi-LSTM model with an F1 value of 68.5 %. [9]

In the paper by Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan have planned LDA based models to interpret the sentiment differences on twitter i.e.-LDA to extract out the foreground topics and RCB-LDA to find out the details why public sentiments have been changed for the target. To determine sentiment, the sentiment analysis tools like SentiStrength and Twitter Sentiment are used whose accuracy is less as compared to other sentiment analysis techniques. [10]

Grounded on the previous research work, mostly sentiment prediction was using Naïve Bayesian and Neural network to categories the text using little amount of data, for the improving the accuracy and treatment the big data deep neural network using construction is applied. In present scenario, applying Machine Learning Methods and Semantic Techniques for handing huge data and improving the accuracy and predicting the right result is the success of the model.

III. PROBLEM IDENTIFICATION

With the speedy growth of the World Wide Web, most of the people are using social media such as Twitter which produces big volumes of opinion texts in the form of tweets that is available for the sentiment analysis. This translates to a vast volume of records from a human perspective which make it difficult to extract a sentence, read them, examines tweet by tweet, review them and organize them into an understandable format in a timely way [13]. Informal language refers to the use of idioms and slang in message, employing the agreements of spoken language such as 'could not' and 'couldn't. Not all systems are able to notice sentiment from use of familiar language and this might hanker the analysis and decision-making procedure

.

IV. THE APPROACH AND METHODOLOGY

Sentiment analysis is a way of defining sentiment of an exact sentence or exact statement. Sentiment analysis is a classification method that develops estimation from the tweets. What other people reflect about us is always a significant part of information throughout the decision making procedure. There are many methods for sentiment analysis. Machine learning methods use numerous machine learning algorithms for classification. Among them Lexicon based method uses dictionary of positive and negative words to identify the sentiment polarity. It is also a based method focus on number of positive words and negative words other than the real meaning. If a tweet has more positive words it shows positive [14]. If a tweet has more negative words it shows negative. In some parts the positive may measure as negative. As the size of dictionary increases new words are separated to positive and negative. The divided words are organized to dictionary order and must be added to words list and added to it becomes hard. To overcome such problems, machine learning methods are used. Machine learning methods are well appropriate for text classification. Such methods learn from earlier calculations to produce dependable decisions and results. Machine learning algorithms like Support Vector machines, random forest, Naïve Bayes, decision trees, and maximum entropy classifiers are used for classification of tweets. Extraction of tweets from twitter is hard. To overcome this problem sentiment analysis of twitter data using machine learning tactics is implemented.

iJETRM
International Journal of Engineering Technology Research & Management

V. MACHINE LEARNING APPROACHES

Sentiment analysis is computerized mining of sentiments, feelings, and emotions from text or speech. It uses machine learning method training set to learn and to train itself. The classifiers are examined by testing data with known inputs and outputs. The stages are shown in figure 1.

Stage1: Collect the data from twitter API
Stage 2: Preprocess the data we collected
Stage 3: Train the models with training set
Stage 4: Test the model
Stage 5: Apply machine learning classification on preprocessed data
Stage 6: See the result in different parameters.

Figure 1. Algorithm for sentiment analysis

1. **Tweet data Collection**: To collect the records from twitter, we need to connect to twitter API and need to allow the twitter API by the access token. To do this with Python use packages Twitter and a 3-Legged OAuth. When the keys are produced, it's much simple to collect the data for a desired product or person etc.

2. **Preprocessing:** The collected records from twitter covers several spelling mistakes, symbols, missing words and urls. Such records pointers to poor results. To avoid this, the preprocessing step is essential to get the correct results. All the collected records are preprocessed in figure 2.

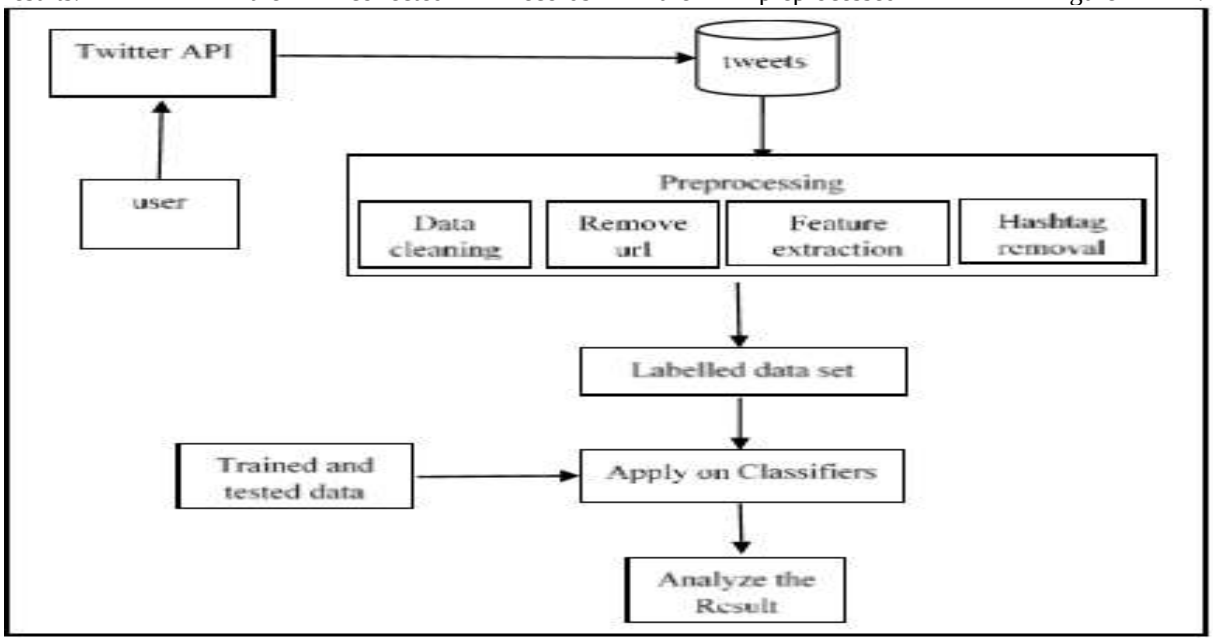

Figure 2. Design for sentiment analysis

This preprocess removes all the urls, special symbols, stop words, hash tags, emoticons etc. this includes syntactical correction of the tweets as needed. The steps included must purpose for making the data machine readable in order to decrease the trouble in nose and feature removal.

5.2.1. **Url Removal** User names and urls current in records are not much significant for the perception of processing. Therefore all the urls and usernames are removed or changed to generic tags.

iJETRM
International Journal of Engineering Technology Research & Management

5.2.2. **Stemming:** It is a process of changing words with their origins, in order to decrease diverse types of words with same meanings. This helps in reduction of dimensionality for nose set.

5.2.3. **Stop Word:** Removal Stop words wish doesn't change the sense of the tweet will be removed or deleted.

5.2.4. **Data Cleaning**: Numbers and special characters existing in tweets doesn't display any sentiment. In certain cases they are diverse with words. Removal of those words helps in link of two words. Otherwise that words are measured as different.

5.2.5. **Feature Extraction**: Once the tweets are cleaned, we need to extract related features for sentiment analysis. The quality and quantity of features is the key for the results generated by a model. This method extracts feature from dataset. After that such features are used to show positive and negative polarity in a sentence which supports in categorizing the view of the people in model. This model extracts the features and procedures the labelled data.

Currently when all the cleaned data is featured and extracted to form a significance full data and this is lastly forms a trained data set which covers positive and negative.

3. **Training and Classification Process:** Supervised learning is a significant method for solving classification problems. In every sentence is early classified as subjective or objective. Only subjective sentences square measure supportive for sentiment classification. Therefore, the goal sentences square measure waste and then the polarity of subjective sentences is calculated. Classifier is applied on the trained data set to train model and the preprocessed record is applied to find the sentiment. Models were applied on the trained data. SVM, Maximum Entropy, Random Forest, Naïve Bayes, models were applied for prediction of sentiment. I will be also analyzing the semantic analysis which was used along with these methods to compute the similarity.

- **Support vector machine (SVM):**

SVM examines the data, describe the decision limitations and uses the kernels for computation which are performed in input space. The input record are two sets of vectors of size m each. Before every records represented as a vector is classified in a specific class. Now the job is to get a margin between two classes that is far from any text. The distance describes the margin of the classifier, maximizing the margin decreases uncertain decisions. SVM also supports regression and classification which are useful for numerical learning model and it helps knowing the factors exactly, that needs to be taken into account, to recognize it positively. SVM precision is 0.89 and the recall is 1.0. The precision is low and recall is high. Most of its predicted labels are not proper.

- **Maximum Entropy:**

Maximum entropy maximizes the entropy distinct on the qualified probability distribution. It level handles overlap nose and is same as logistic regression that gets distribution over classes. It follows positive feature exclusion limits.

$$P_{ME}(c|d,) = \frac{\exp[\Sigma_i \lambda_i f_i\ (c,d)]}{\Sigma_{c'} \exp\ [\Sigma_i \lambda_i f_i\ (c,d)]}$$

Where, c is the class, d is the tweet, and is a weight vector. The weight vectors choose the significance of a feature in classification. It follows the parallel processes as naïve bayes, discussed above and delivers the polarity of the sentiments. The precision for Maximum Entropy is 0.93 and recall is 1. The precision is low and recall is high. Most of its predicted labels are not proper.

- **Naive Bayes:**

The naïve bayes has been used since of its simplicity and easiest in both through training and classifying step. It is a probabilistic classifier that can learn the pattern of examining a set of documents which has been categorized.

iJETRM
International Journal of Engineering Technology Research & Management

$$P_{NB}(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)}$$

This compare the contents with the list of words to categorize the documents to their right group. The class c* is given to tweet d, where, the f signifies a feature and ni (d) signifies the count of feature fi found in tweet d. where we can see that there are a total of m features. Parameters P(c) and P (f|c) are found through maximum probability guesses which are incremented by one for smoothing. The pre-processed record along with removed feature is provided as input for training the classifier using naïve bayes. When the training is complete, during classification this provides the polarity of the sentiments. Naïve Bayes precision is 0.85 and recall is 0.7. Precision is high and recall is low. That means most of its predicted labels are correct.

- **Random Forests:**

The random forests are a joint learning method for classification which completes with the help of building a large number of selection trees at training time and outputting the class that is process for the training output through individual trees. The random forest produces multi-altitude selection trees at input phase and output is produced inside the form of multiple decision trees. The connection among trees is reduced by randomly deciding on trees and as a result the prediction strength will increase and results to successful in performance. Predictions are made by combining the predictions of various joint information units. The precision is 1.0 and recall is 0.8. Here precision is high and recall is low. That means most of its predicted labels are correct.

- **Decision Tree:**

The decision tree classifier is a supervised learning algorithm that can be used for both classification and regression tasks. It can be modified almost too several type of records. This splits the training data into small parts in order to classify outlines so that they can be used for classification. Such algorithm is definitely used where there are many hierarchical categorical differences can be made. The decision tree algorithm contains of root node, decision node and leaf node. Where the root node represents the entire data set and decision node performs computation and leaf node produces the classification. In training phase, this algorithm studies what are the decisions that are to be made in order to split the labelled data into its classes. Passing the records through tree, an unknown case is classified. The calculation that proceeds place in each of the decision node usually matches the selected feature with determined constant, the decision will be made based on whether the nose is greater or less than the constant by creating two way split in the tree. The data will be eventually passed through these decision nodes until that spreads a leaf node which represent its assigned class. The precision is 1 and recall is 1. Both precision and accuracy are high which results in all labels are correct for the decision tree.

- **Semantic Analysis:**

Semantic analysis is resulting from the WordNet database where every term is related with every other. This database is of English words that are linked together. Supposed two words are close to each other, they are semantically similar. More exactly, I will able to control synonym like parallel. I am going map terms and examine their relationship in the ontology. The key mission is to use the stored documents that cover terms and then check the similarity with the words that the user uses in their sentences. Thus it is supportive to show the polarity of the sentiment for the users.

iJETRM
International Journal of Engineering Technology Research & Management

Name of model	Precision	Recall	F-Score
SVM	0.89	1.0	0.9
Maximum Entropy	0.93	1.0	0.96
Naïve Bayes	0.85	0.7	0.76
Random Forest	1.0	0.8	0.8
Decision Tree	1.0	1.0	1.0

Table 3. Precision, Recall and F-score values

Among all of them Decision tree is having best f-score after they are compared to all models shown in table 3.

VI. RESULT AND CONCLUSION

The present work focuses on sentiment analysis using machine learning approaches. The approach needs training by using a dataset from collected data. The methods can be applied to unknown data after training. The methods like Naïve Bayes, Support Vector Machines, Decision Tree, Random Forest and Maximum Entropy are implemented. The Naïve Bayes algorithm is mainly focus and the result is compared with other models in terms of precision, recall and F-Score. Among all these, decision tree scores the highest value. I have implemented different well-known deep learning models for twitter sentiment analysis and preprocessed data to remove noise from data and increase the accuracy of models.

REFERENCES

[1] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very deep convolutional networks for text classification,” arXiv preprint arXiv:1606.01781, 2016.
[2] S. M. Mohammad, S. Kiritchenko, and X. Zhu, “Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets,” arXiv preprint arXiv:1308.6242, 2013.
[3]M. Bouazizi and T. Ohtsuki, “Multi-class sentiment analysis on twitter: Classification performance and challenges,” Big Data Mining and Analytics, vol. 2, no. 3, pp. 181–194, 2019.
[4] Vidhya Content Team. 2015. Quick Guide: Steps To Perform Text Data Cleaning in Python [ONLINE] Available at: https://www.analyticsvidhya.com/blog/2015/06/quick-guide-textdatacleanin Goodfellow-et-al-2016, [Accessed 20 May 2017].
[5] 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.
[6] Zhang, X. and Zheng, X., 2016, July. Comparison of text sentiment analysis based on machine learning. In 2016 15th International Symposium on Parallel and Distributed Computing (ISPDC) (pp. 230 233). IEEE
[7] R. Joshi and R. Tekchandani, "Comparative analysis of Twitter data using supervised classifiers," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-6. doi: 10.1109/INVENTIVE.2016.7830089
[8] 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.

iJETRM
International Journal of Engineering Technology Research & Management

[9] J. Barnes, R. Klinger, and S. S. i. Walde, "Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets," arXiv preprint arXiv: 1709.04219, 2017.
[10] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, "Interpreting the Public Sentiment Variations on Twitter", IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO.5, MAY 2014.
[11] D. Reynard, M. Shirgaokar, "Harnessing the power of machine learning: Can Twitter data be useful in guiding resource allocation decisions during a natural disaster?", Transportation Research Part D: Transport and Environment, 2019.
[12] K. Ghazvini, M. Yousefi, F. Firoozeh, S. Mansouri, "Predictors of tuberculosis: Application of a logistic regression model", Gene Reports, Vol(17), 2019.
[13] Y. Armani, M. Lazaar, K. Kadiri, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis", Procedia Computer Science, Vol(127), 511–520, 2018.
[14] 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.
[15] L.Ramachandran,E.F.Gehringer, "Automated Assessment of Review Quality Using Latent Semantic Analysis," ICALT, IEEE Computer Society, pp. 136-138, 2011.
[16] Ding, X.; Liu, B.; Yu, P.S. A holistic lexicon-based approach to opinion mining. In Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, CA, USA, 11–12 February 2008; pp. 231–240.