IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 12, Issue 10, October 2022, ||Series -I|| PP 36-44

Sentiment Analysis in Twitter Data utilizing Supervised Machine Learning

G VIJAYA KUMARI[1], B S R D LAKSHMI[2]
[1]*Associate Professor, Department of Computer Science Engineering, BVCITS, AMALAPURAM, AP*
[2]*Assistant Professor, Department of Computer Science and Engineering, BVCITS, AMALAPURAM, AP, INDIA.*
Received 29 October 2022; Accepted 20 November 2022

Abstract:
Sentiment analysis is an extremely important procedure these days for social Network analysis. Sentiment analysis or assessment mining is the procedure of consequently separating data from notions or suppositions of others about some subject or issue. We can recognize suppositions in a substantial unstructured/organized data and investigate extremity of sentiments. Twitter is a vast and quickly becoming small scale blogging long range informal communication site where individuals express their assessments in short and straightforward way of expressions. Tweets can be broke down to perform conclusion analysis on different substances. In this paper we concentrate on various methodologies utilized for sentiment analysis of twitter data. Informal communication destinations like Twitter, Facebook, Google+ are quickly picking up fame as they permit individuals to share and express their perspectives about points, have talk with various groups, or post messages over the world. In this paper, we give a study and a relative analysis's of existing methods for sentiment mining like machine learning and vocabulary based methodologies, together with assessment measurements. Utilizing different machine learning algorithm like fuzzy, we give research on twitter data streams. We have likewise talked about general difficulties and utilizations of Sentiment Analysis on Twitter.
Keywords: Twitter, Sentiment analysis (SA), Social media, post.

I. Introduction

Microblogging today has turned into an exceptionally mainstream specialized instrument among Internet users. A great many messages are seeming every day in well-known sites that give administrations to microblogging, for example, Twitter, Tumblr, and Facebook. Writers of those messages expound on their life, offer suppositions on assortment of themes and examine current issues. On account of a free configuration of messages and a simple availability of microblogging stages, Internet users tend to move from customary specialized devices, (for example, conventional online journals or mailing records) to microblogging administrations. As more users post about items and administrations they utilize, or express their political and religious perspectives, microblogging sites get to be important wellsprings of individuals' sentiments and sentiments. Such data can be effectively utilized for promoting or social studies. We utilize a dataset framed of gathered messages from Twitter. Twitter contains an expansive number of short messages made by the users of this microblogging stage. The substance of the messages fluctuate from individual considerations to open articulations. Table 1 indicates case of normal posts from Twitter. As the crowd of microblogging stages and administrations develops each day, data from these sources can be utilized as a part of conclusion mining and estimation analysis assignments. For instance, fabricating organizations might be keen on the accompanying inquiries: • what do individuals consider our item (benefit, organization and so on)? • How constructive (or adverse) are individuals about our item? • What might individuals favor our item to resemble? Political gatherings might be intrigued to know whether individuals bolster their project or not. Social associations may ask individuals' sentiment on current level headed discussions. This data can be gotten from microblogging administrations, as their users post ordinary what they like/aversion, and their assessments on numerous parts of their life. In our paper, we concentrate how microblogging can be utilized for conclusion analysis purposes. We demonstrate to utilize Twitter as a corpus for conclusion analysis and supposition mining. We utilize microblogging and all the more especially Twitter for the accompanying reasons: • Microblogging stages are utilized by various individuals to express their assessment about various themes, in this way it is a profitable wellspring of individuals' sentiments. • Twitter contains a huge number of content posts and it develops each day. The gathered corpus can be self-assertively huge. • Twitter's gathering of people fluctuates from general users to big names, organization agents, legislators, and even nation presidents. In this manner, it is conceivable to gather content posts of users from various social and interests bunches. • Twitter's gathering of people is spoken to by users from numerous nations. In spite of the fact that users from U.S. are winning, it is conceivable to gather data in various dialects. We gathered a corpus of 300000 content posts from Twitter uniformly split naturally between three arrangements of writings: 1. writings containing positive sentiments, for example, joy, delight or happiness 2. Writings containing negative sentiments, for example, misery, indignation or frustration 3. Target messages that exclusive express a reality or don't express any sentiments we play out a phonetic analysis of our corpus and we demonstrate to assemble a slant classifier that uses the gathered corpus as preparing data.

II. Related Work

There are two procedures generally used to recognize the sentiments from content. They are Symbolic methods and Machine Learning systems [3]. A. Estimation analysis utilizing Symbolic Techniques A typical system utilizes the accessibility of lexical assets. Turney [4] recommended a methodology for notion analysis called 'sack of words'. In the specified methodology, singular words are disregarded and just accumulations of words are considered. He assembled word having descriptive words or qualifier for the extremity of audit from a web index Altavista. A lexical database called WordNet [6] was utilized by Kamps et al [5] which decides an enthusiastic matter in a word. WordNet conveys equivalent words and separation metric to discover the introduction of descriptors. To defeat deterrents in lexical substitution undertaking, Baroni et al [7] built up a framework bolstered by word space model formalism along these lines speaking to nearby words. EmotiNet theoretically spoke to the content that put away the structure of genuine occasions in a space. This was presented by Balahur et al [8]. B. Assessment analysis utilizing Machine Learning Techniques Under this method, there are two sets, in particular a preparation set and a test set. For the most part the dataset which is gathered from various sources and whose conduct and yield qualities are known not falls into the classification of preparing data sets. Conversely with this, the datasets whose qualities or conduct are obscure to us are called as test data sets. Here various classifiers are prepared with preparing data and after that obscure data or we can say a test data is given to this model to get craved results. Machine Learning comprises of different diverse classifiers, for example, Ensemble classifier, k-implies, Artificial Neural Network and so on. These are utilized to order audits [8]. Y.Mejova et al [1] in his exploration work suggested that we can utilize nearness of every character, recurrence of events of every character, word which is considered as invalidation and so forth as elements for making highlight vector. He additionally demonstrates that we can adequately utilize unigram and bigram ways to deal with make highlight vector in Sentiment analysis. Domingos et al [10] recommended that Naive Bayes functions admirably for ward highlights for certain issue. Zhen Niu et al [11] found another model. This model depends on Bayesian calculation. In this model, some proficient methodologies are utilized for selecting highlight, calculation of weight and grouping. Barbosa et al [12] planned a 2 stage analysis technique which is a programmed estimation analysis for ordering tweets. In the initial step, tweets are ordered into subjective and target tweets. After that, in a brief moment step, subjective tweets are delegated positive and negative tweets. Celikyilmaz et al [13] created one strategy as articulation based word bunching. This technique standardizes uproarious tweets. There are a few words which have the same articulation yet having distinctive implications. Along these lines, for disposing of this contention, there is strategy specified previously. In this said technique, words having same articulation are bunched and appointed basic tokens. Wu et al [14] in his paper prescribed model, specifically, the impact likelihood to analysis the sentiment tweets. In this, if @username is found in the tweet, it makes affecting move and serves to impacting likelihood. By gathering programmed tweets, Pak et al [15] built up a technique for slant analysis by making twitter corpus. In his proposed work he demonstrates that, while making highlight vector, we can utilize emoticons as a component. He utilized a Naïve Bayesian classifier to do the sentiment analysis. Some investigates made to distinguish the popular supposition about motion pictures, news and so forth from twitter tweets. V.M. Kiran et al [16] had taken the data from other openly accessible databases like IMDB and Blippr.

III. Methodology

The proposed method for sentiment analysis in this paper could be represented in 5 stages, each of which are listed below:

A. Data Collection
B. Data Preprocessing
C. Feature Selection
D. Model Selection
E. Model Evaluation

A. Data Collection

Data collection is the first phase for analysis as there needs to be data for us to do analysis on. In our experimentations we have used python programming language as a tool. Being that said, data collection in this particular analysis could be carried out in two ways. First way is to collect preorganized data from different sites such as kraggle . On these sites this preorganized data is uploaded by the developers of sites themselves or is posted by different researchers for free. All one needs to do to acquire this data is to create a free account on these sites. Second way is to manually extract data from twitter using some API available for twitter. For this we have chosen tweepy as an API for extraction of tweets. Tweepy does not compatible with the new versions of python(python 3.7) . So for using this particular API an older version of python is needed(python 2.7). To access tweets on twitter using API first we need to authenticate the console from which we are trying to access twitter. This could be done by following steps listed below:

- Creation of a twitter account.
- Logging in at the developer portal of twitter.
- Select "New App" at developer portal.
- A form for creation of new app appears, fill it out Fill.
- After this the app for which the form was filled out will go for review by twitter team.
- Once the review is complete and the registered app is authorized then and only then the user is provided with

'API key' and 'API secret'

- After this "Access token" and "Access token secret" are given.

These keys and tokens are unique for each user and only with the help of these can one access the tweets directly form twitter. For this paper we have extracted a large data set consisting of almost 3000 tweets. These tweets are taken using #USairlines and thus are about different US Airlines. We have used textblob package of python for predata annotation of polarity for these tweets.

Data set	No. of tweets
Training data	2343
Testing data	585

Table 1:- Data Distribution\

B. Data Preprocessing

The pre-processing of data implies the processing of raw data into a more convenient format which could be fed to a classifier in order to better the accuracy of the classifier. Here, in our case the raw data which is being extracted from twitter using an API is initially totally unstructured and bogus as the availability of various useless characters seems very common in it.

For this matter we remove all the unnecessary characters and words from this data using a module in python known as Regular Expressions, are for short. This module adopts symbolic techniques to represent different noise in the data and therefore makes it easy to drop them. Specifically in twitter terminology there are various common useless phrases and spelling mistakes present in the data, which need to be removed to boost the accuracy of our resultant. These could be summoned up as follows:

- Hash tags: these are very common in tweets. Hash tags represent a topic of interest about which the tweet is being written. Hashtags look something like #topic.
- @Usernames: these represent the user mentions in a tweet. Sometimes a tweet is written and then is associated with some twitter user, for this purpose these are used.
- Retweets(RT): as the name suggests retweets are used when a tweet is posted twice by same or different user.
- Emoticons: these are very commonly found in the tweets. Using punctuations facial expressions are formed in order to represent the smile or other expressions, these are known as emoticons.
- Stop words: stop words are those word which are useless when it comes to sentiment analysis. Words such as it, is, the etc are known as stop words.

C. Feature Selection

As mentioned earlier in this paper different researchers have used different features for the classification of the tweets, in our experimentations similar feature are being used. These features include Unigram, Bigram, N-gram, POS tagging, Subjective, objective features and so on. NLTK short for Natural Language Tool Kit is another module available in python which also open source and could be used for extraction of these features.

D. Model Selection

Once the data is being pre-processed, this data is to be fed to a classification model for further processing. There are different classification algorithms on which these models are built on. In this paper, we have chosen k-nearest neighbour model to perform the classification.

KNN or k-Nearest Neighbour algorithm represents a machine learning technique used for classifying a set of data into its given target values (in our case positive , neutral or negative).KNN could also be used for regression problems but is widely used for classification problems.

Now, any classification model needs a target set on which we train the model for its further use. As for mentions in the literature survey section most of them have manually set these target values to positive, negative or null. For this paper we have used a library in python known as Textblob to automatically set the target for each tweet. The data set then is divided into two halfs training set and testing set. The data set used by us in our experimentations consisted of 2928 tweets so we segregated it into training and testing data. Training portion consisted of 2343 tweets whereas the test set consisted of 585 tweets. Now this training as well as test set needs to be transformed into binary values so as to be fed to the model. The models don't understand any values other than the binary.

For this we have used another module of the python known as sklearn which contains many classification model as well as different encoders in it. For this paper this library is being for model selection, label encoding as well as model evaluation which would be mentioned in next section.

E. Model Evaluation

One of the most common and appropriate technique used for evaluation of a classifier is through confusion matrix. A generalized form of confusion matrix is given in table 4.5 below:

	Predicted class1	Predicted class 2
Actual class 1	True positive(tp)	False negative(fn)
Actual class 2	False positive(fp)	True negative(tn)

Table 2:- General Confusion Matrix

By applying this technique we can derive the generalized evaluation parameters. These parameters include:

- *Accuracy* : accuracy of a classifier indicates how accurately the classifier has predicted the result.It can be calculated using the formula:

$$accuracy(\mathrm{a}) = \frac{tp + tn}{tp + tn + fp + fn}$$

- *Precision*: precision shows how often the result that is being predicted by the classifier, when it indicates true is correct. The formula for precision is:

$$precision(p) = \frac{tp}{tp + fp}$$

- *Recall*: it indicates the true positive rate of the classifier. Formula for recall is:

$$recall(r) = \frac{tp}{tp + fp}$$

- *F1 score :* it indicates the weighed average of recall and precision.Formula for recall is:

$$F1\ score = \frac{2p.r}{p - r}$$

IV. Proposed Work

The success of both this learning methods is mainly depends on the selection and extraction of the specific set of features used to detect sentiment. The machine learning approach applicable to sentiment analysis mainly belongs to supervised classification.

In a machine learning techniques, two sets of data are needed:

1. Training Set
2. Test Set.

A number of machine learning techniques have been formulated to classify the tweets into classes. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and fuzzy (FUZZY) have achieved great success in sentiment analysis. Machine learning starts with collecting training dataset. Nextly we train a classifier on the training data. Once a supervised classification technique is selected, an important decision to make is to select feature. They can tell us how documents are represented.

The most commonly used features in sentiment classification are

- Term presence and their frequency
- Part of speech data
- Negations
- Opinion words and phrases

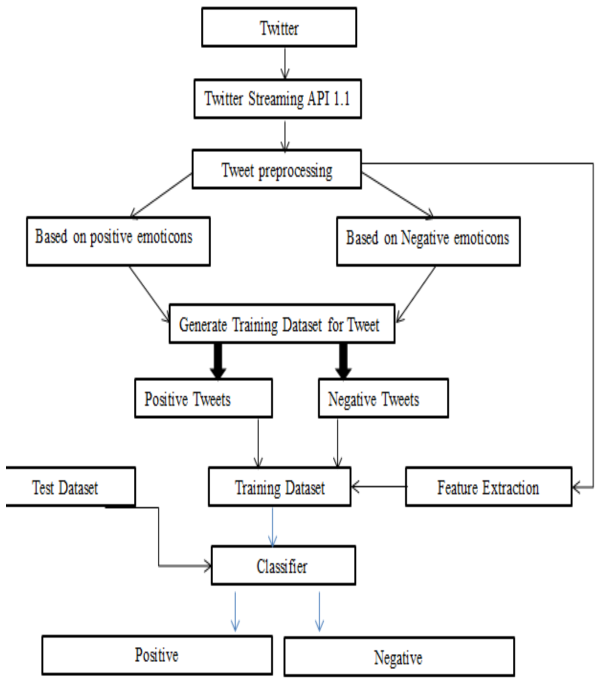

Fig 1. Proposed System Architecture

With respect to unsupervised techniques, fuzzy (FUZZY), Maximum Entropy are some of the most common techniques used. Whereas semi-supervised and unsupervised techniques are proposed when it is not possible to have an initial set of labeled documents/opinions to classify the rest of items

V. Details of Modeling

Initially, we acquire a dataset and partition the dataset into training and test data. The major approach used to classify the tweets is the machine learning technique. The tweets may be rife with slang words and misspellings. So, we need to perform sentence-level analysis for all these tweets. This is done in three phases. In the first phase, pre-processing is done. This is done mainly to eliminate the slang words, misspellings and other faults. In the second phase, a fuzzy is created using relevant features. Finally, using different classifiers, we will be able to classify the tweets as positive, negative or neutral.

A) Acquiring the dataset

We use the Sanders analytics dataset to perform necessary actions. Sanders analytics dataset consists of a total of 5600 tweets containing tweets of companies like Apple, Google and Microsoft. The dataset is labeled and therefore, we know exactly which tweets are positive, negative, neutral and irrelevant.

B) Pre-processing tweets

Keyword extraction is difficult in Twitter due to misspellings and slang words. So to avoid this, a pre-processing step is performed before feature extraction. Pre-processing steps include removing URLs, avoiding misspellings and slang words. Misspellings are avoided by replacing repeated characters with 2 occurrences. Slang words contribute considerably to the emotion of a tweet. So they can't be simply removed. Therefore, a slang word dictionary is maintained to replace slang words occurring in tweets with their associated meanings. Domain data contributes much to the formation of slang word dictionary. Also, we use a technique in which if the overall sentiment of the tweet is obtained, we will be able to find out the sentiment score of the new term by just looking at its relative position in the sentence.

C) Creation of fuzzy

Feature extraction is done in two steps. In the first step, Twitter-specific features are extracted. Hashtags and emoticons are the relevant Twitter-specific features. Emoticons can be positive or negative. Therefore, they are assigned different weights. Positive emoticons are assigned a weight of '1' and negative emoticons are assigned a weight of '-1'. There may be positive and negative hashtags. Therefore, the count of positive hashtags and negative hashtags are added as two separate features in the fuzzy. Twitter-specific features may not be present in all tweets. So, a further feature extraction is to be done to obtain other features. After extracting Twitter-specific features, they are removed from the tweets. A tweet can be then treated as simple text. Thereafter, using unigram approach, tweets are represented as a collection of words. In unigrams, a tweet is represented by its keywords. We maintain a negative keyword list, positive keyword list and a list of different words that represent

negation. Counts of positive and negative keywords in tweets are used as two different features in the fuzzy. Presence of negation contribute much to the sentiment. So their presence is also added as a relevant feature.
All keywords cannot be treated equally in the presence of multiple positive and negative keywords. Therefore, a special keyword is selected from all the tweets. In the case of tweets having only positive keywords or only negative keywords, a search is done to identify a keyword having relevant part of speech. A relevant part of speech is an adjective, an adverb or a verb.Such a relevant part of speech is defined, based on their relevance in determining sentiment**.** A keyword that is an adjective, adverb or a verb shows more emotion than others. If a relevant part of speech can be determined for a keyword, then it is taken as special keyword. Else, a keyword is selected randomly from the available keywords as special keyword. If both positive and negative keywords are present in a tweet, we select any keyword having relevant part of speech. If relevant part of speech is present for both positive and negative keywords, none of them is chosen. Special keyword feature is given a weight of ’1’ if it is positive and ’-1’ if it is negative and ’0’ in its absence. Part of speech feature is given a value of ’1’ if it is relevant and ’0’ otherwise.
Thus, fuzzy is composed of 8 relevant features. The 8 features used are part of speech (pos) tag, special keyword, presence of negation, emoticon, number of positive keywords, number of negative keywords, number of positive hashtags and number of negative hashtags.

d) Sentiment analysis

After creating the fuzzy, sentiment classification is done using a combination of both knowledge-based and machine learning approach. Word-by-word sentiment analysis using knowledge-based approach is used along with different classifier techniques using fuzzy.

Computing new word Sentiment

We derive day-to-day sentiment scores by counting positive and negative messages. Positive and negative words are defined by the subjectivity lexicon from OpinionFinder, a word list containing about 1,600 and 1,200 words marked as positive and negative respectively (Wilson, Wiebe, and Hoffmann 2005). We do not use the lexicon’s distinctions between weak and strong words.
A message is defined as positive if it contains any positive word, and negative if it contains any negative word. (This allows for messages to be both positive and negative.) This gives similar results as simply counting positive and negative words on a given day, since Twitter messages are short.
A major issue when using such short messages is the presence of misspellings, emoticons, links and other unnecessary content. The pre-processing stage helps us remove these shortcomings considerably and makes the resultant text clean. In order to compute the sentiment score for a new word, we simply work the other way around. Once we have the score of the text, we simply associate the words with the general feel of the entire text. So, if soccer comes between great and happy we would associate soccer with a positive sentiment score.

Evaluation

Data contains 4685 rows. We create a subset of data by removing the irrelevant factor altogether. Resulting data set contains 3126 objects containing 5 variables. The recipe book contains the following variables:

1) Topic
2) Tweet id
3) Tweet Text
4) Tweet date
5) Sentiment

Topic: Apple, Microsoft, Google. Tweets concerning Apple, Microsoft and Google are used in the analysis of sentiment score of tweets.
Sentiment: This column actually describes the sentiment class of each tweet. Possible values are positive, negative, irrelevant and neutral.
Tweet ID: Contains the Twitter ID of the tweet
Tweet Date: Contains the date of the tweet
Tweet Text: Contains the actual text of the tweet
The original Sanders analytics data set histogram is shown below:

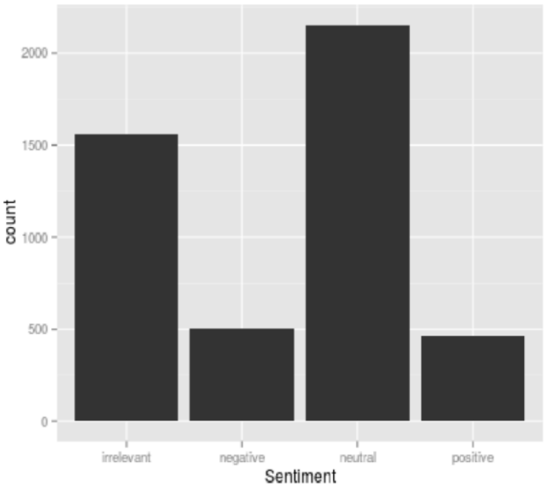

Fig 2. Histogram before eliminating irrelevant tweets

On performing a summary operation on the resultant data set, we get the following table.

Table 1. Tweet classification before elimination of irrelevant tweets

Irrelevant	Negative	Neutral	Positive
1559	509	2153	464

As we can see evident from the histogram, the dataset contains irrelevant tweets. Therefore, we eliminate them from the dataset. After eliminating the irrelevant tweets, the histogram obtained as follows:

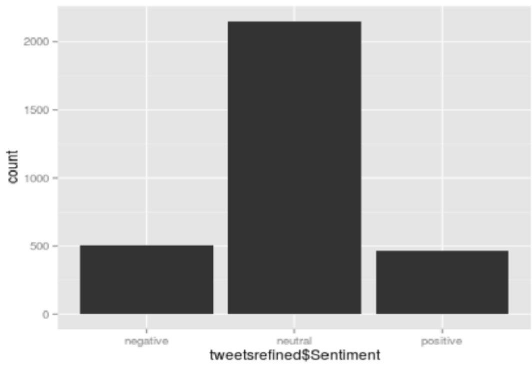

Fig 3. Histogram after eliminating irrelevant tweets

On performing a summary operation after the elimination of irrelevant tweets from the data set, we get the following table.

Table 2. Tweet classification after elimination of irrelevant tweets

Irrelevant	Negative	Neutral	Positive
0	509	2153	464

The following table demonstrates the scores that were assigned to 10 random tweets from the dataset.
Class: The label that was assigned to the tweet in the dataset
Tweet: The tweet text
Score: The sentiment score calculated by our model

Table 3. Sentiment score of 10 tweets using hybrid approach

Class	Tweet	Score
Positive	Now all @Apple has to do is get swype on the iphone and it will be crack. Iphone that is	2
Positive	@Apple will be adding more carrier support to the iPhone 4S (just announced)	3
Positive	Lmao I think @apple is onto something magical! I am DYING!!! haha. Siri suggested where to find food and where to hide a body lolol	3
Positive	Currently learning Mandarin for my upcoming trip to Hong Kong. I gotta hand it to @Apple iPhones & their uber useful flashcard apps	1
Negative	Why is #Siri always down @apple	-1
Negative	I just need to exchange a cord at the apple store why do I have to wait for a genius? @apple	-5
Negative	@apple AirDrop #fail - Immediate "declined your request." every time	-8
Negative	Dear @apple My new Air is now a notbook since your update killed #wifi #bug #destroying #productivity	-3
Neutral	Using @Apple's mobile @AirPort Utility http://t.co/TiDpaHYC	0
Neutral	#motoactiv? Methinks @apple and maybe @Nike are already prepping lawsuits	0

	Irrelevant	Negative	Neutral	Positive
Irrelevant	0	0	0	0
Negative	0	509	0	0
Neutral	0	0	2153	0
Positive	0	0	0	464

Thus, as we can see, we get an accuracy of 100%.

After training the classifier using hybrid method which combines knowledge-based approach and machine learning capabilities we get the following confusion matrix.

VI. Conclusion

Thus we conclude that the machine learning technique is very easier and efficient than symbolic techniques. These techniques are easily applied to twitter sentiment analysis. Twitter sentiment analysis is difficult because it is very tough to identify emotional words form tweets and also due to the presence of the repeated characters, slang words, white spaces, misspellings etc. To handle these problems the fuzzy is created. Before creating fuzzy pre-processing is done on each tweet. Then features are extracted in two phases: First phase is the extraction of the twitter specific word. Then they are removed from the text. Now extracted fuzzy is transformed into normal text. After that, features are extracted from tweet which is normal text without any hash tags or slang words. And these extracted features are then added to form fuzzy. There are different machine learning classifiers to classify the tweets. From our results, we have shown that Naïve Bayesian and Support vector machine performs well and also provide higher accuracy. The results show that we get 75 % accuracy form FUZZY and 65% accuracy form Naïve Bayesian classifier. So we can increase the accuracy of classification as we increase the training data. By this project we can say that fuzzy performs better for tweets related to Movie reviews.

REFERENCES

[1]. Twitter as a Corpus for Sentiment Analysis and Opinion Mining Alexander Pak, Patrick Paroubek.

[2]. T.K Das, D.P Acharjya, M.R Patra, " Opinion mining about a product by analysing public tweets in twitter," IEEE Proceedings of International Conference on Computer Communication and Informatics (ICCI-2014), Jan 03-05, 2014, Coimbatore, India

[3]. A. Shrivatava, S. Mayor and B. Pant, "Opinion Mining of Real Twitter Tweets," International Journal of Computer Applications, Volume 100- No. 19, August 2014.

[4]. Anton Barhan, Andrey Shakhomirov, "Methods for Sentiment Analysis of Twitter Tweets," Proc. 12th Conference of Frucst Association (2012).

[5]. Shailesh Kumar Yadav, "Sentiment Analysis and Classification: A survey", International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 3, March 2015..

[6]. Dilara Totungolu, Gurkan Telseren, Ozgun Sagturk & Murat C. Ganiz, "Wikipedia Based Semantic Smoothing For Twitter Sentiment Classification", IEEE (2013).

[7]. Chihli Hung, Hao- Kai Lin, "Using Objective Words in SentiWordNet to Improve Sentiment Classification for Word of Mouth", IEEE 2013.

[8]. Antonio Moreno-Ortiz, Chantal Pere Hernandez, "Lexicon- Based Sentiment Analysis of Twitter Messages in Spanish," ISSN 1135- 5948, pp 93-100, 2013.

[9]. A. Khan, B. Baharudin, K.Khan,"Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure" ICSECS 2011: 2nd International Conference on Software Engineering and Computer Systems, Springer, pp. 317-331, 2011s.

[10]. L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis,"Technical report, HP lLaboratories, 2011.

G VIJAYA KUMARI, et. al. "Sentiment Analysis in Twitter Data utilizing Supervised Machine Learning." *IOSR Journal of Engineering (IOSRJEN),* 12(10), 2022, pp. 36-44.