A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data

Priyanka Tyagi[1] and Dr. R.C. Tripathi[2]

Abstract—Any opinion of an individual through which the feelings, attitudes and thoughts can be expressed is known as sentiment. The kinds of data analysis which is attained from the news reports, user reviews, social media updates or microblogging sites is called sentiment analysis which is also known as opinion mining. The reviews of individuals towards certain events, brands, product or company can be known through sentiment analysis. The responses of general public are collected and improvised by researchers to perform evaluations.

The popularity of sentiment analysis is growing today since the numbers of views being shared by people on the microblogging sites are also increasing. All the sentiments can be categorized into three different categories called positive, negative and neutral. Twitter, being the most popular microblogging site, is used to collect the data to perform analysis. Tweepy is used to extract the source data from Twitter. Python language is used in this research to implement the classification algorithm on the collected data.

The features are extracted using N-gram modeling technique. The sentiments are categorized among positive, negative and neutral using a supervised machine learning algorithm known as K-Nearest Neighbor.

Keywords: Sentiment Analysis, Classification Techniques, Literature Review, Future Scope

I. INTRODUCTION

A. Sentiment Analysis

Over the past few years, an interesting and popular research area emerging lately is sentiment analysis. The opinions that are held by any numbers of individuals are reviewed and analyzed using sentiment analysis [1]. These reviews can be related to an event, brand, person or product. Earlier, magazines, newspapers and other sources were used to express people views. However, with the advancement in technology the people have began to express their feelings on different social networking and microblogging sites. In a productive manner, the opinions of individuals have been extracted, studied and then evaluated by researchers. Twitter has gained the highest popularity in comparison to all other microblogging platforms in the past few years. It can be considered as a valid indicator for the sentiments of people. Different ways have been developed by several media organizations to mine the twitter information[1]

- conduct training, testing and analysis, the tweets are collected using API.
- *Topics:* On any imaginable topic the messages are posted by the Twitter users. This is different from other microblogging sites in which only particular topic and purpose is discussed.
- *Real time:* Since the blogs are longer and huge amount of time needs to be invested, these blogs are updated at longer time intervals [4].
- Few basic terminologies related to twitter are used by the users. Some of these terminologies are described below:

B. Data Extraction

The data source was collected from the tweets posted on Twitter. The twitter API is used to extract the tweets from Twitter. The “twitteroauth” version of the public API is used and implemented in PHP. Either the web servers or local hosts can execute this directly or for the query few parameters are considered. An extensive set of filtering parameters are set during the extraction of tweets from Twitter such that they can match any specific criteria. API is used to keep the query running after it has been generated. The output of this query will be all the relevant twitter source data. The data is directly embedded within the MySQL database for being utilized in future. In every record that is generated though a tweet, information such as tweet id, text, client name and so on can be extracted. If any user makes his location public, the data relevant to the location from where tweet is posted is generated in the form of latitude and longitude from the Twitter API. However, because of the security issues and client protections, people have stopped sharing their locations

[1]Research Scholar, Department of Mgt, LDIMS,
Mandi Road New Delhi, India
[2] Professor, Department of CSE,
Lingaya’s Vidyapeeth Faridabad, Haryana, India
E-mail: [1]priya21.tyagi@gmail.com
[2]rctripathi@lingayasuniversity.edu.in

since 2012. The location is used as a filtering parameter in the principle query later on Twitter. Thus, depending upon the settled set of locations, the tweets are extracted [6].

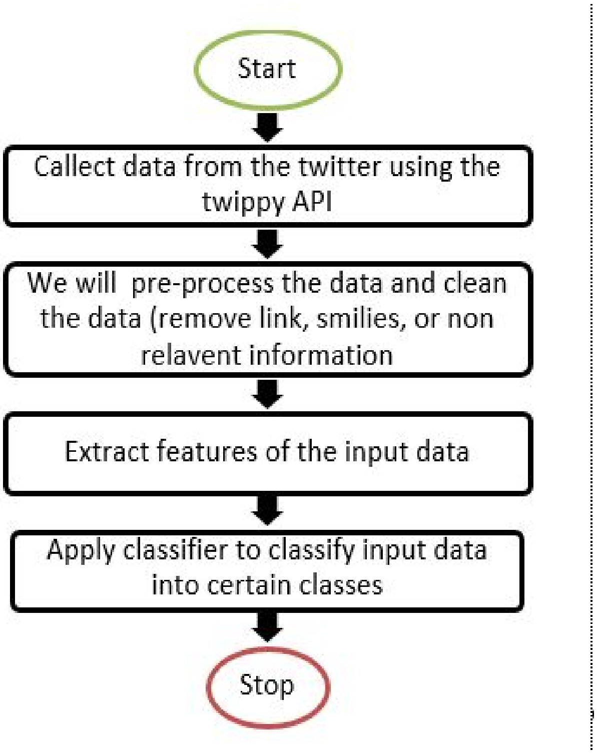
Start
Callect data from the twitter using the twippy API
We will pre-process the data and clean the data (remove link, smilies, or non relavent information
Extract features of the input data
Apply classifier to classify input data into certain classes
Stop

Figure 1: Flow chart for sentiment analysis

C. Data Preprocessing

There is certain amount of irrelevant data available within the data that is extracted from Twitter. Any kinds of arbitrary characters or useless information need to be filtered out from the tweet information. The Natural Language Processing tool is applied for filtering out this useless data. Any kinds of grammatical relations that exist in between the words of sentences are given as output by this NLP tool. Within the general natural language research it is not useful to include certain advanced linguistics available in English language. Thus, there are 50 predefined relations called dependencies available in NLP which are listed and explained in this standard description [7]. The information analysts consider the main word relations as important even though linguistics defines a few other word relations within a sentence, due to which these 50 dependencies have been defined in NLP. The most used dependencies amongst these 50 are nsubj, amod, dobj. The tweets that contain meaningful information are recognized using these relations. The outcomes are not helped at any rate through the facilitating filtering along with more relations. The relations among nouns and adjectives or verbs are discovered using nsubj relation within any noun sentence. Irrespective to complementing a noun in a sentence or not, this is considered to be of high importance.

D. Data analysis

1) Creation of Dataset

The Twitter API is used to extract the dataset. Positive and negative classes are created for classifying the tweets. By including around 600 positive and 600 negative tweets, a dataset is generated here [8].

2) Preprocessing of Tweets

The extraction of keywords becomes difficult due to the presence of slangs and incorrect spellings in tweets. Thus, a preprocessing step is performed for filtering out the slang words and misspellings before extracting the features. Any slang words present in the tweets are replaced with their relevant meanings using the slang word dictionary. The slang word dictionary is created using the domain information.

3) Creation of Feature Vector

The features are extracted from tweets in the next step. In the initial step, specific features like hash tags and emoticons are extracted. On the basis of polarity of emotions they depict, particular weights are assigned to the emoticons on the basis of their polarity. The positive emoticons are assigned with weight “1” and negative emoticons with “-1” weight. It is possible for hash tag to be positive and negative. Within the vector of features, they are included as individual features.

4) Sentiment Classification

A feature vector is designed using standard classifiers such as SVM, ensemble classifiers, Naïve Bayes classifiers and so on within the classification step.

E. Classification Techniques

A text classification issue needed to be resolved is sentiment analysis. Machine learning approach and Lexicon based approach are the two broader categorizations of these classification approaches.

F. Machine Learning Approaches

The text is classified into classes by the machine learning based approach using classification techniques. The two broader categorizations of these machine learning techniques are:

Unsupervised learning: There is no category involved and the targets are not provided by them at all. Thus, clustering is considered to be an important factor here.

Supervised learning: The labeled dataset is used to develop this method. When the classification approach is to be designed, the labels are provided to the model. For getting significant outputs when going through decision making these labeled datasets are trained [11].

The determination and extraction of particular sets of features such that the sentiments can be detected is the success of both of these learning techniques.

Naive Bayes (NB), Maximum Entropy (ME), and Support Vector machines (SVM) are few amongst the widely used machine learning techniques for sentiment classification. When having an initial set of labeled opinions is unrealistic for training the classifier, the semi-supervised and un-supervised techniques are designed.

II. Naive Bayes Classifier

The considerable numbers of features are utilized in feature vector through Naïve Bayes classifier [12]. Since these features are independent equally, analyzing them exclusively is important. The mathematical representation of conditional probability for Naïve Bayes is given as:

$$P(X|yj) = \prod_{i=1}^{m} P(x_i|y_i)$$

A feature vector denoted by "X" is included here which is defined by X={x_1,x_2,....x_m}. The class label is represented by yj. The classification of different types of independent features such as positive and negative keywords, emoticons and emotional keywords is done efficiently using Naïve Bayes. The relationships amongst features are not considered in Naïve Bayes classifier. Thus, the relationships which exist among emotional keyword, negation words and speech tag are not utilized in it.

III. Support Vector Machine Classifier

Huge margin is used for classification through SVM classifier. A hyper plane is used to differentiate the tweets. A discriminative function is utilized by SVM as:

$$gX=wT\varphi X+b$$

The feature vector is denoted in the above equation by "X", weights vector by "w" and the bias vector by "b". The non-linear mapping which transforms information space to high dimensional feature space is denoted by φ(). On the training set, w" and "b" are recognized automatically. A linear kernel is applied for classification in this approach [13].

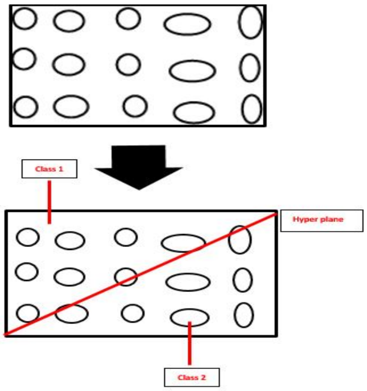

Figure 2: SVM Classifier

IV. Maximum Entropy Classifier

With respect to the relationship amongst features, no assumptions are considered in maximum entropy classifier. The conditional distribution of class label is estimated by maximizing the entropy of system through this classifier. The mathematical representation of conditional distribution is:

$$P\lambda(y|X) = 1/Z(X)\exp(\sum_{i} \lambda_i f_i(X,y))$$

Here, the feature vector is represented by "X" and the class label by "y". The normalization factor is represented by Z(X) and the weight coefficient by λi. For classification within our feature vector, the relationships amongst part of speech tag, emotional keyword and negation are utilized.

V. Ensemble Classifier

Different kinds of ensemble classifiers are developed. For performing the best classification, all the features of all the best classifiers are utilized in this classifier. Naive Bayes, Maximum entropy and SVM are the three approaches utilized by the base classifiers. The voting rule is used to create an ensemble classifier. Depending upon the output of larger parts of classifiers, their classification is done.

Following are the two sets of data needed within machine learning approaches [14]:

a. Training Set

A. *Test Set*

The training dataset is collected to initiate machine learning. The training data is used in the next step for training a classifier. Selecting the feature is an imperative decision to be made after the selection of a supervised classification approach. The representation of documents can be known through this. During sentiment classification, the most commonly used features include:

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

When having an initial set of labeled opinions is unrealistic for training the classifier, the semi-supervised and unsupervised techniques are designed.

The sentiment dictionary which consists of opinion words is used by lexicon based approach. The polarity is determined by matching these words with rest of the data. For understanding how positive, negative and objective the words contained in dictionary are, the sentiment scores are assigned to opinion words. The sentiment lexicon which is an accumulation of known and precompiled sentiment phrases, idioms and terms is used as a base for the lexicon-based approaches. For different traditional genres of communication, this approach is developed [15].

This approach has two sub classifications:

B. *Dictionary-based*

The terms which are collected normally and then annotated in a manual way are utilized for this approach. The synonyms and antonyms of a particular word within the dictionary are searched for growing this set. WordNet is an example of one such dictionary using which a thesaurus called SentiWordNet is developed. The domain and context based orientations cannot be managed by this method which is its major drawback.

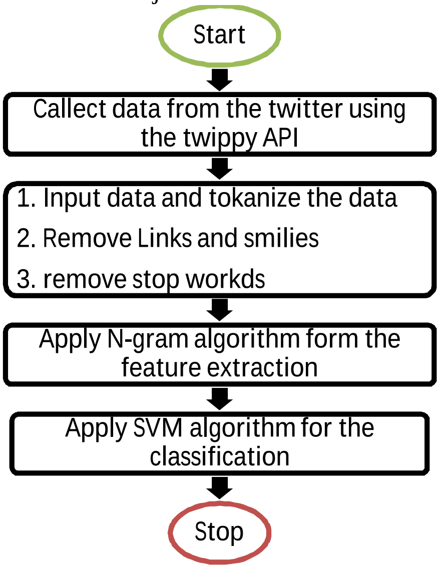
Start
Callect data from the twitter using the twippy API
1. Input data and tokanize the data
2. Remove Links and smilies
3. remove stop workds
Apply N-gram algorithm form the feature extraction
Apply SVM algorithm for the classification
Stop

Figure 3: Detailed flow Chart of sentiment analysis

C. *Corpus-Based*

The dictionaries related to a particular domain are provided by the corpus-based approach. A set of seed opinion terms which grow from the search of relevant words using statistical or semantic techniques are created by these dictionaries. Following are the two basic methods that are based on statistics [16]:

- Latent Semantic Analysis (LSA).
- An interesting solution can be provided by different methods which are based on semantics.

A comparative investigation of existing opinion mining techniques that include cross domain, machine learning, and lexicon based techniques and so on has been provided depending upon certain performance measures such as recall and precision.

VI. Literature Survey

A study related to automatic text summarization which is gaining popularity lately. For covering all the important contents and general information, a compressed version of documents is created. Few features are used to score sentences within extractive text summarization. In past few studies, large numbers of features network based techniques are proposed. In order to score the sentences, each of the features which use metrics and idea of complex network have been reviewed [17]. Discussion of experimental results on single component and combinations of various features are made. The assessments being performed on DUe 2002 data sets include the quantitative and qualitative aspects. For summarization shortest ways were provided using which the highest scores for increasing the quality were achieved. The results that were achieved by integrating similar kinds of network properties were another contribution of this approach. The sentences were chosen on the basis of this incredible influence. The two categorizations of text summarization techniques are extractive and abstractive methods. This paper presented a comprehensive survey of both of these techniques used for text summarization [19]. This paper studied the different summarization techniques. An effective summary that has less redundancy and includes grammatically correct sentences is to be generated through the summarization approach. The users can use extractive and abstractive methods from which efficient results are achieved. For generating compressed and readable information for users, the hybridization technique proposed here proves to be highly efficient as per the test results. Mihai Dascălu, *et.al*, (2011) proposed an automatic approach with respect to NLP and used distributing figuring to improve the runtime performance [25]. Further, with respect to the multilayered engineering an exceptional grading component is provided. A replicated worker design is sent to improve the speed of this process. In this approach, along with increment in performance level, there are two important aspects to be considered which are, load balancing and fault tolerance. The corpus of chats can be assessed in a timely manner resulting is providing a quick access to the participants' feedbacks as per this demonstration of framework. By investigating huge corpuses in very less time, a solution is provided through which the general performance is improved. This is done by utilizing the subtle elements on a distributed form of instrument. Under different conditions and loads, the performance of proposed work is better as compared to existing approaches.

VII. Problem Formulation

The behaviors of users are analyzed through the sentiment analysis technique. There are different social network websites on which sentiment analysis is applied. The pattern matching algorithm is applied in sentiment analysis to extract the features of input data. Further, the classification techniques are applied in order to detect sarcasm. The N-gram approach was used in the technique proposed in base paper for extracting the features from social networking sites. Along with neural networks, the pattern-matching was applied in this previously proposed technique. For the classification of features, naïve bayes classifier was applied. There were two major issues highlighted in this research. As per the first issue, the colored features were only extracted from social networking sites through N-gram algorithm. However, it is important to analyze the texture features as well to increase the efficiency of this technique. The SVM classifier is used in the existing approach for the sentiment analysis. The SVM classifier only classify data into two classes and also it has accuracy approximately of 80 percent which reduce efficiency of sentiment analysis.

VIII. Conclusion

This research is based on analyzing the sentiments of users using sentiment analysis technique. Twitter data is used to perform sentiment analysis. The tokenization is performed on twitter data and sentiment analysis is performed by calculating the polarity of data. The sentiment analysis is performed in base paper technique using SVM classifier. To perform sentiment analysis four steps are performed which are pre-processing, tokenization, feature extraction and classification. The technique of SVM is used in the existing work for the classification. The SVM classifier gives accuracy of approx 80 percent for the sentiment analysis. To increase accuracy of sentiment analysis, the hybrid classification approach will be designed in this research work based on naïve bayes and KNN. The proposed approach will be implemented in python and it is expected that accuracy will be increase for the sentiment analysis of twitter data.

References

[1] O. Rambow, L. Shrestha, J. Chen, and C. Lauridsen. *Summarizing email threads.* In Proceedings of HLT-NAACL 2004: Short Papers, pages pp.105–108, 2004.

[2] G. Salton and C. Buckley. *Term-weighting approaches in automatic text retrieval. Information Processing and Management,* 24: pp. 513–523, 1988.

[3] O. Sandu. *Domain Adaptation for Summarizing Conversations.* PhD thesis, Department of Computer Science, The University Of British Columbia, Vancouver, Canada, 2011.

[4] S. Teufel and M. Moens. *Summarizing scientific articles: Experiments with relevance and rhetorical status.* Computational Linguistics, 28: 409–445, 2002.

[5] J. Ulrich, G. Murray, and G. Carenini. *A publicly available annotated corpus for supervised email summarization.* In AAAI08 EMAIL Workshop, Chicago, USA, 2008. AAAI.

[6] D.C. Uthus and D. W. Aha. *Plans toward automated chat summarization.* In Meeting of the Association for Computational Linguistics, pp.1–7, 2011.

[7] C. Whitelaw, B. Hutchinson, G. Chung, and G. Ellis. *Using the web for language independent spellchecking and autocorrection.*In Empirical Methods in Natural Language Processing, pp. 890–899, 2009

[8] L. Zhou and E. H. Hovy. *Digesting virtual geek culture: The summarization of technical internet relay chats.* In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 298–305, 2005

[9] Tharindu Weerasooriya, Nandula Perera, S.R. Liyanage. *A method to extract essential keywords from tweet using NLP.* 2016 16th International Conference on Advances in ICT for Emerging Regions (ICTer).

[10] Ibrahim A. Hameed. *Using Natural language processing for designing socially intelligent robots.* 2016 Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob).

[11] L. SUanmali, M. S. Binwahlan, and N. Salim. *Sentence features fusion for text summarization using fuzzy logic in Hybrid Intelligent Systems.* 2009, HIS'09, Ninth International Conference on, vol. 1, IEEE, 2009, pp. 142–146.

[12] L. Suanmali, N. Salim, and M.S. Binwahlan. *Fuzzy logic based method for improving text summarization.* arXivpre print arXiv:0906.4690, 2009.

[13] X.W. Meng Wang and C. Xu. *An approach to concept oriented text summarization,* Proceedings of ISClTS05, IEEE international conference, China, pp. 1290–1293" 2005.

[14] M.G. Ozsoy, F.N. Alpaslan, and 1. Cicekli. *Text summarization using latent semantic analysis.* Journal of Information Science, vol. 37, no. 4, pp. 405–417, 2011.

[15] Adyan Marendra Ramadhani, Hong Soon Goo. *Twitter Sentiment Analysis using Deep Learning Methods.* 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 2017.

[16] K. Kaviya, C. Roshini, V. Vaidhehi, J. Dhalia Sweetlin. *Sentiment for Restaurant Rating.* 2017 IEEE International Conference on Smart Technologies and Management for Computing, Controls, Energy and Material (ICSTM).

[17] Dan Cao, Liutong Xu. *Analysis of Complex Network Methods for Extractive Automatic Text Summarization.*2016 2nd IEEE International Conference on Computer and Communications.

[18] RasimAlguliyev, Ramiz Aliguliyev, NijatIsazade. *A Sentence Selection Model and HLO Algorithm for Extractive Text Summarization.* 2016, IEEE.

[19] Narendra Andhale, L.A. Bewoor. *An Overview of Text Summarization Techniques.* 2016, IEEE.

[20] RupalBhargavaandYashvardhan Sharma. *MSATS: Multilingual Sentiment Analysis via Text Summarization.* 2017, IEEE.

[21] Archana N. Gulati, Dr. S.D. Sawarkar. *A novel technique for multi-document Hindi text summarization.* 2017 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017).

[22] Manisha Gupta, Dr.Naresh Kumar Garg. *Text Summarization of Hindi Documents using Rule Based Approach.*2016 International Conference on Micro-Electronics and Telecommunication Engineering.

[23] Akshi Kumar, Aditi Sharma, SidhantSharma,Shashwat Kashyap. *Performance Analysis of Keyword Extraction Algorithms Assessing Extractive Text Summarization.* International Conference on Computer, Communication, and Electronics (Comptelix), 2017.

[24] N. Moratanch, S. Chitrakala. *A Survey on Extractive Text Summarization.* IEEE International Conference on Computer, Communication and Signal Processing (ICCCSP), 2017.

[25] Mihai Dascălu, CiprianDobre, Ştefan Trăuşan-Matu, Valentin Cristea. *Beyond Traditional NLP: A Distributed Solution for Optimizing Chat Processing.* 2011 10th International Symposium on Parallel and Distributed Computing.

[26] Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. *Understand Short Texts by Harvesting and Analyzing Semantic Knowledge.* 2016, IEEE.

[27] Pierre Ficamos;Yan Liu, WeiyiChenA. *Naive Bayes and Maximum Entropy approach to sentiment analysis: Capturing domain-specific data in Weibo.* 2017 IEEE International Conference on Big Data and Smart Computing (BigComp).

[28] Ankur Goel, Jyoti Gautam, Sitesh Kumar. *Real time sentiment analysis of tweets using Naive Bayes.* 2016 2nd International Conference on Next Generation Computing Technologies (NGCT).

[29] Shweta Rana, Archana Singh. *Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques.* 2016 2nd International Conference on Next Generation Computing Technologies (NGCT).

[30] Huma Parveen, Shikha Pandey. *Sentiment analysis on Twitter Data-set using Naive Bayes algorithm.* 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).

[31] Wiraj Udara Wickramaarachchi, R.K.A.R. Kariapper. *An Approach to Get Overall Emotion from Comment Text towards a Certain Image Uploaded to Social Network Using Latent Semantic Analysis.* 2017 2nd International Conference on Image, Vision and Computing.