International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

A Novel Approach for Hateful Speech Detection of Twitter Dataset: An Overview

Priyanka Ramgir[1], Neha Ovhale[2,] Madhuri Shivade[3], Nikita Manjalkar[4], Prof. D.D. Ahir[5]

[1-4] UG Students, Modern Education Society's College of Engineering, Pune, India

[5]Assistance Professor, Modern Education Society's College of Engineering, Pune, India

ABSTRACT: Hate speech is of current and broad interest in the domain ofMedia Networks. The Internet has both anonymity and flexibilityhas made it easy for users to interact aggressively. And asThere is an increasing amount of online hate speech, methods which automaticallyit's much needed to identify hate speech. These problems also haveThe Natural Language Processing and Machine Learning societies were also attracting considerable. Hence the aim of this paper is to analyze howthe application of natural language contributes in the identification of hate speech. In addition, this paper also applies an existing technique on a dataset in this field.. The results from the case study are quite encouraging. In the proposed research we proposed a sentiment classification approach using Recurrent Neural Network (RNN) and Naïve Bayes algorithm. The twitter base streaming dataset has used for classification. Natural Language Processing and Machine Learning algorithms has used for classification for training as well as testing respectively. The performance evaluation shows how proposed system is better than classical classification algorithms.

KEYWORDS: NLP, Machine Learning, Twitter data, ANN, NB.

I.INTRODUCTION

In conjunction with data mining, classical methods rely on the manual feature engineering and rules. The manual configuration of data functionsInstances in vectors of the features can be made in several ways. Analysis hasShowed that the most powerful surface features when detecting hate speechBag n-grams of words, words and characters. As for the classifiers, the mostThe Support Vector Machine is common algorithm used. Much like algorithmsThe classification task also includes Naive Bayes, Logistic Regression, and Random Forest.In the past decade, new forms of communication, such as micro blogging and text messaging have emerged and become ubiquitous. While there is no limit to the range of information conveyed by tweets and texts, often these short messages are used to share opinions and sentiments that people have about what is going on in the world around them. Tweets and texts are short: a sentence or a headline rather than a document. Another aspect of social media data such as Twitter messages is that it includes rich structured information about the individuals involved in the communication. There is a growing number of people who hold accounts on social media platforms (SMPs) but hide their identity for malicious purposes. Sentiment Analysis is process of computationally identifying and categorizing opinions expressed in a piece of text to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative. Sentiment analysis is to extract the opinion of the user from the text document.In the proposed work we find and analyze the issue of Bag-of-Words model as it disrupts the word order an d discards some of semantic structure, and Familiar problem in Polarity Shift Problem during Negation of Sentiment. The general practice in sentiment classification follows the techniques in traditional topic-based text classification

II.LITERATURE SURVEY

An Approach to Detect Abusive Bangla Text[1]Our goal is to detect abusive Bangla comments which are collected from various social sites where people share their sentiment, opinions, views etc. in this paper. We proposed a root level algorithm to detect abusive text and also proposed unigram string features to get a better result.

Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection [2].An approach to detect hate expressions on Twitter. Our approach is based on unigrams and patterns that are automatically collected from

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

~the training set. These patterns and unigrams are later used, among others, as features to train a machine learning algorithm. Our experiments on a test set composed of 2010 tweets show that our approach reaches an accuracy equal to 87.4% on detecting whether a tweet is offensive or not (binary classification), and an accuracy equal to 78.4% on detecting whether a tweet is hateful, offensive, or clean (ternary classification).
Research on text sentiment analysis based on CNNs and SVM[3].A Convolutional Neural Networks (CNNs) model combined with SVM text sentiment analysis is proposed. The experimental results show that the proposed method improves the accuracy of text sentiment classification effectively compared with traditional CNN, and confirms the effectiveness of sentiment analysis based on CNNs and SVM.

An approach AHTDT- Automatic Hate Text Detection Techniques in Social Media[4].The automatic hate text detection techniques and highlights the parametric comparison of those techniques. Different social networking platforms like YouTube, Facebook, Whisper, Blogger etc. contains different hate text detection techniques which depend on natural language processing, data mining and machine learning domains. Detecting hate text in social media provides an online safety to youngsters. Techniques based on Bag of Word (BOW) approaches have a few constraints such as; a BOW model ignores the semantics of the word.
A framework for sentiment analysis with opinion mining of hotel reviews[5].A framework is termed sentiment polarity that automatically prepares a sentiment dataset for training and testing to extract unbiased opinions of hotel services from reviews. A comparative analysis was established with Naïve Bayes multinomial, sequential minimal optimization, compliment Naïve Bayes and Composite hypercube on iterated random projections to discover a suitable machine learning algorithm for the classification component of the framework. Irecent years, Twitter has become one of the most popularmicro-blogging social-media platforms, providing a platformfor millions of people to share their daily opinions/thoughts usingreal-time status updates Conover et al. (2013). Twitter has 270Million active users and 500 million tweets are sent per day.

Consequentially, we need to rule out certain conditions when identifying hate speech. For example, if we try to explain the meaning of some abusive words or if we use some of the racial terms in another context that has no undertone of hate. Add to this when writing a news article and referring to a "hate crime" sect. This referral of ISIS itself will not be considered a speech of hate. Similarly, Waseem and Hovy[6] suggested 11 criteria for identifying hate speech directly on a twitter forum, some of which are: the use of sexiest and ethnic words, targeting and insulting minorities, encouraging violence, distorting the facts with lies and endorsing suspicious hashtag.

Given these characteristics, a reasonable list with certain adjustments can be derived for a particular culture to deal with the controversy and then from that list, hate speech can be reliably identified and acknowledged. Anis[7] discussed the dominant themes in Arabic hate speech particularly in the newspaper, and concluded that hate speech in the Arab region is usually related to religion and sectarian themes.

Hate speech is difficult to grasp. It can however be recognized based on specific features that can be distinguished from one culture to another. Such features are debatable, they may be viewed by some as mere hate and by some as not. This problem is considered a controversial issue that nobody can come to terms with. Gelashvili and Nowak[8] argued that it is an obstacle for owners of social media platforms to regulate hate speech, as many questions like what constitutes hate speech are raised to their heads? And what kind of hate speech does it take to counteract? Only legitimate individuals who are actively engaged in the same culture and who can be sufficiently competent can provide answers to those questions. Some studies have given some necessary terminologies to study hate speech, for example Fortuna and Nunes[9] have listed some of the key rules for identification of hate speech. In short, when dismissing prejudice about community, hate speech is defined. Along with the use of racial and sexist slurs intended to harm. Add to that when talking indecently about a particular country or religion.

Natural Language Processing or (NLP) is the main pillar of text mining, employing a number of computational tasks to trace and understand human natural language by the machine [10]. Today, by downloading huge amounts of unstructured data, NLP researchers have moved towards the rich and controversial data available in social networks, these data can be mined and put into practical use. Text mining for social networks requires a number of lexical, syntactic and semantic NLP tasks aimed at giving the text a structure to be processed further. These tasks include: tokenisation, which breaks the text by the spaces into word tokens. Also, avoid words like "in," "the" will be taken out of this job, since they don't make sense of the context. There are many tokenization tools available, such as "OpenNLP1" or "Stanford Tokenizer," Apache. Then, in

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

order to provide lexical information, predicting Part of Speech (PoS) for each token must take place. Thereafter, parsing takes place by representing the syntactic structure of the entire text[11]. A major drawback of NLP nowadays is that most of the tools are designed exclusively for common languages such as English, French, and Spanish [10]. In comparison, uncommon language such as Arabic has a challenge associated with the difficulty of adapting the tools of common languages. However, specialists in Arabic linguistics have obtained considerably good results in studying morphology of the Arabic language.

III.SYSTEM DESIGN

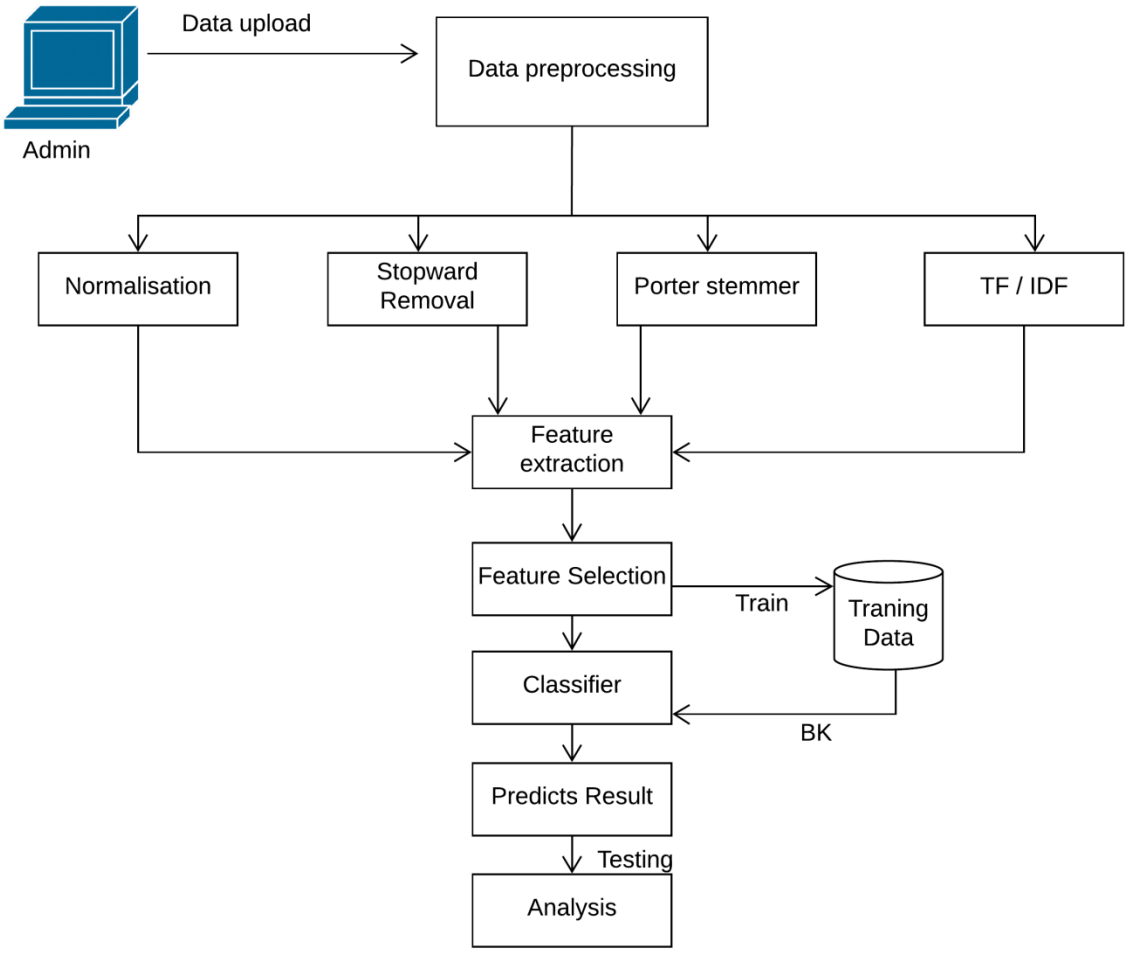

Figure 1 : Proposed System Architecture

For this research work system collect the data from various internet sources like twitter API and some synthetic sources. The above architecture illustrated in figure 1 which describes first system download the data set according to input queries using Twitter API. Many times in internet data contains some miss classified instances, it must to need preprocess the data as well as normalization. Natural language processing (NLP) is the initial step of preprocessing which contains tokenization, stopword removal, porter stemmer and TF-IDF respectively. Once the all processes has done system extract the respective features from entire data set. The unigram method has used extract feature. Once feature extraction has done to select best feature from available data. Cosine similarity with TF-IDF, correlation with TF-IDF methods used to select the best feature. The hybrid method for feature selection which is used for training as well as testing.Finalysystem execute the classification algorithm to create the background knowledge in train model, which is the level for positive as well as negative. In the testing phase system execute NLP process for entire data set to extract and select the feature set. Artificial

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

Neural Network (ANN) and Naive Bayes (NB) classifiers has used to predict the accuracy. Finally in results section system shows performance evaluation of system with existing classification algorithms.

Algorithms Used

1 : Stop word Removal Approach

Input: Stop words list L[], String Data D for remove the stop words.

Output: Verified data D with removal all stop words.

Step 1: Initialize the data string S[].

Step 2: initialize a=0,k=0

Step 3: for each(read a to L)

If(a.equals(L[i]))

Then Remove S[k]

End for

Step 4: add S to D.

Step 5: End Procedure

2 Stemming Algorithm.

Input : Word w

Output : w with removing past participles as well.

Step 1: Initialize w

Step 2: Intialize all steps of Porter stemmer

Step 3: for each (Char ch from w)

If(ch.count==w.length()) && (ch.equals(e))

Remove chfrom(w)

Step 4:if(ch.endswith(ed))

Remove 'ed' from(w)

Step 5: k=w.length()

If(k (char) to k-3 .equals(tion))

Replace w with te.

Step 6: end procedure

3 TF-IDF

Input : Each word from vector as Term T, All vectors V[i…n]

Output : TF-IDF weight for each T

Step 1 : Vector = {c1, c2, c3….cn}

Step 2 : Aspects available in each comment

Step 3 : D = {cmt1, cmt2, cmt3, cmtn}

and comments available in each document

Calculate the Tf score as

Step 4 :tf (t,d) = (t,d)

t=specific term

d= specific document in a term is to be found.

Step 5 :idf = t → sum(d)

Step 6: Return tf *idf

4. Artificial Neural Network

Input : Training Rules Tr[], Test Instances Ts[], Threshold T.

Output : Weight w=0.0

Step 1 : Read each test instance from (TsInstnace from Ts)

Step 2 : TsIns $= \sum_{k=0}^{n}\{Ak \ldots An\}$

Step 3 : Read each train instance from (TrInstnace from Tr)

Step 4 :$TrIns = \sum_{j=0}^{n}\{Aj \ldots \ldots Am\}$

Step 5 :*w = WeightCalc(TsIns, TrIns)*

Step 6 :if (w >= T)

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

Step 7 : Forward feed layer to input layer for feedback FeedLayer[] ← {Tsf,w}
Step 8 :optimized feed layer weight, Cweigt←FeedLayer[0]
Step 9 : Return Cweight

IV.RESULTS AND DISCUSSIONS

The proposed system performance evaluation, we calculate matrices for accuracy. We implement the system on java architecture framework with INTEL 3.0 GHz i5 processor and 8 GB RAM. Some user comments are positive or negative, and the data contains around 80,000 user's comments. The system finally classifies all the comments as positive, negative as well as neutral. Negation handling also works at the time of aspect classification. Here table 1 shows the estimated system performance with different existing systems. So, proposed results are around on satisfactory level.

Table 1: Performance Analysis of Proposed System

Approach	Feature selection	Data Source	Accuracy
Lexical Resource	POS Apriori	Amazons customers Reviews	87.07%
Lexical Approach	Graph Distance Measurement	Users Blog Posts	82.85%
Hybrid	n-gram	Movie based review	90.05%
Naive Bayes	Information Gain	Canteen services reviews	91.75%
Naive Bayes and SVM	Based on minimum cuts	Movie reviews from users	85.90%
Proposed Approach	NLP and ML	Specific Product based Review	95.90%

V.CONCLUSION

The volume of the real data has several implications. The volume imposes the demand that the methods used to process this data be sufficiently fast as slower models may not be able to process on the incoming data in real time. On the other hand, it allows us to filter the data and to concern ourselves only with the portion of the data that meets a certain level of quality. In proposed research the preprocessing module contains the sub modules for tokenization, Stop word removal and Stemming. Synonym handling is done for aspects with same meaning but different name to avoid ambiguity. It was e.g. observed that not all of the posts bearing negative sentiment contain a reasonable complaint, expressing the cause of the negative sentiment. The sentiment class itself will in most cases be not the final product of a practical application. The product should contain useful information that can be acted upon. Therefore, we should restrain ourselves to those social network posts that are clearly related to a product, service or an aspect of the former ones. The posts then should be aggregated in order to present the results in an easily comprehensible manner. This should be one of the aims of focus of further research.Furthermore, accuracy can be increased in future by enhancing features set and testing for other classification techniques such as deep learning with different activation functions. The performance of system can be increased by using other techniques such as Deep Learning with different activation functions in future.

REFERENCES

[1] Watanabe, Hajime, MondherBouazizi, and TomoakiOhtsuki. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection." IEEE Access 6 (2018): 13825-13835.
[2] Gaydhani, Aditya, et al. "Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach." arXiv preprint arXiv:1809.08651 (2018).
[3] Wiegand, Michael, et al. "A survey on the role of negation in sentiment analysis." Proceedings of the workshop on negation and speculation in natural language processing. 2010.
[4] Al-Hassan, Areej, and Hmood Al-Dossari. "Detection of hate speech in social networks: asurvey on multilingual corpus." Computer Science & Information Technology (CS & IT) 9.2 (2019): 83.
[5] Badjatiya, Pinkesh, et al. "Deep learning for hate speech detection in tweets." Proceedings of the 26th International Conference on World Wide Web Companion.International World Wide Web Conferences Steering Committee, 2017.
[6] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People?Predictive Features for Hate Speech Detection on Twitter," Proc. NAACL Student Res. Work., pp. 88–93, 2016.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

[7] M. Y. Anis and U. S. Maret, "Hatespeech in Arabic Language," in International Conference on Media Studies, 2017, no.September.
[8] T. Gelashvili and K. A. Nowak, "Hate Speech on Social Media," Lund University, 2018.
[9] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Comput.Surv., vol. 51, no. 4, pp. 1–30, 2018.
[10] J. Hirschberg and C. D. Manning, "Advances in natural language processing," Science (80-.)., vol. 349, no. 6245, p. 261 LP-266, Jul. 2015.
[11] S. Sun, C. Luo, and J. Chen, A review of natural language processing techniques for opinion mining systems, vol. 36, no. November 2016. 2017.