International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

Hateful Speech Detection form short text using NLP and Machine Learning Techniques

Vaishali Vanjari[1], Shreyas Gaikwad[2] , Rutuja Saste[3], Satyam Indhane[4], Prof. B.C.Julme[5]

BE Students, Department of Computer Engineering, Pune Vidyarthi Grih's College of Engineering & Technology, Pune, Maharashtra, India [1,2,3,4]

Professor, Department of Computer Engineering, Pune Vidyarthi Grih's College of Engineering & Technology, Pune, Maharashtra, India[5]

ABSTRACT: Social media constitute a rich source of information. The analysis of the big volume of user-generated content in social media can provide meaningful information for gleaning people's emotions and deeper understanding public attitude and mood. In this work, we present a generic framework for the analysis of big social data and the modeling of public emotions and mood. The framework consists of two main parts. Initially an ensemble classifier schema, which combines a generic domain independent, knowledge-based tool with machine learning methods, is used to recognize emotional content in user generated social data. Then, a graph based method is utilized to model the emotional level of a topic based on the emotions recognized and then it creates the topic's emotional graph visualizing public emotions and mood on the topic. The results from the case study are quite encouraging. In the proposed research we proposed a sentiment classification approach using Support Vector Machine (SVM), Decision Tree (DT) and Naïve Bayes (NB) algorithm. The twitter base streaming dataset has used for classification. Natural Language Processing (NLP) and Machine Learning (ML) 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, SVM, NB, Decision Tree

I.INTRODUCTION

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 and 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.but, Problem of this technique are 1) Not handling a word or phrase is not accurate. 2) Not handling the polarity shift problem.The proposed system plans to solve this problem by using methods like Natural language processing, machine learning, etc. we propose a simple yet efficient model, called Naïve bayes, SVM and DT etc.

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
~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.

Title	Authors	Year	Methodology	Gap Analysis
Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection [10]	HAJIME WATANABE, MONDHER BOUAZIZI and TOMOAKI OHTSUKI,	2017	Proposed approach automatically detects hate speech patterns and most common unigrams and usesthese along with sentimental and semantic features to classify tweets into hateful, offensive and clean. Our proposed approach reaches accuracy equal to 87.4% for the binary classification of tweets into offensive and non-offensive, and accuracy equal to 78.4% for the ternary classification of tweets into, hateful, offensive and clean.	Failed to build a richer dictionary of hate speech patterns that can be used, along with a unigram dictionary, to detect hateful and offensive online texts.
Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach,	AdityaGaydhani, Vikrant Domay, ShrikantKendre and LaxmiBhagwat	2018 [1]	Perform experiments considering n-grams as features and passing their term frequency-inverse document frequency (TFIDF) values to multiple machine learning models	The model does not account for negative words present in a sentence. Improvements can be done in this area by incorporating linguistic features
The model does not account for negative words	IlhamMaulana Ahmad	2018	Machine learning can recognize any kind of hate speech on the image through the existing text. With using Latent Semantic Analysis	Less accuracy as compared to previous work

present in a sentence. Improvements can be done in this area by incorporating linguistic features,	Niam, BudhiIrawan, CasiSetianingsih, BagasPrakoso Putra		(LSA) method, we get the result of this research is precision 67%, recall 76.84%, and accuracy 57.9%.	
Detection of hate speech in social network : A survey [5]	Areej Al-HassanandHmood Al-Dossari,	2019	Present a background on hate speech and its related detection approaches. In addition, the recent contributions on hate speech and its related anti-social behaviour topics will be reviewed. Finally, challenges and recommendations for the Arabic hate speech detection problem will be presented	Failed include incorporating the latest deep learning architectures to build a model that is capable to detect and classify Arabic hate speech in twitter into distinct classes.
Deep Learning for Hate Speech Detection in Tweets, [6]	PinkeshBadjatiya, Shashank Gupta, Manish Gupta andVasudevaVarma,	2017	Extensive experiments with multiple deep learning architectures to learn semantic word embedding's to handle this complexity. Our experiments on a benchmark dataset of 16K annotated tweets show that such deep learning methods outperform state-of-the-art char/word n-gram methods by ~18 F1 points.	Failed to explore the importance of the user network features for the task.
Hate me, hate me not: Hate speech detection on Facebook, [7]	Fabio Del Vigna, Andrea Cimino, FeliceDell'Orletta, Marinella Petrocchi and Maurizio Tesconi	2017	Design and implement two classifiers for the Italian language, based on different learning algorithms: the first based on Support Vector Machines (SVM) and the second on a particular Recurrent Neural Network named Long Short Term Memory (LSTM)	Refinement of the classifier results is required.
Hate Speech Detection on Indonesian Instagram Comments using FastText Approach, [4]	Nur Indah Pratiwi, Indra Budi, and IkaAlfina,	2018	FastText as the classifier and word representation. The experiment results showed that FastText is better than Random Forest Decision Tree and Logistic Regression. The highest result achieved when FastText is combined with bigram feature with F-measure of 65.7%.	The size of the dataset used in this study is very small. In the next work we suggest to build bigger dataset; To build dictionary of offensive words/phrases for Indonesian language as suspect that offensive words often occurred inhate speech comments
Automated detection of hate speech towards woman on twitter[8]	Havvanur Sahi, YasimenKilic and Rahime	2018	A supervised learning model which is developed to classify cyber hate towards on twitter Turkish twits, with Hashtag specific to choice of clothing for woman, have been collected and five machine learning based classification algorithms were applied including SVM (using polynomial and RBF Kernel), J48, Naïve Bayes, Random Forest	Failed to focus on those hard to classify instances and try sentiment analysis to improve results

	Belen Saglam		and Random Tree.	
Analysis Text of Hate Speech Detection Using Recurrent Neural Network, [9]	Arum SuciaSaksesIMuhammadNasrun, CasiSetiaingsih,	2018	By using Deep Learning method with Recurrent Neural Network (RNN) algorithm. After the creation of this program, it is hoped the computer can know and classify the existence of hate speech in the sentence. From the results of tests thathave been done the average precision of 91%, recall 90% and accuracy 91%	The more training data the better the resulting Accuracy.
Hate Speech on Twitter A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection, , [2]	Hajime Watanabe, Mondher Bouazizi, Tomoaki Ohtsuki	2018	Approach is based on unigrams and patterns that are automatically collected from the training set. These patterns and unigrams are later used, among others, as features to train a machine learning algorithm	Failed to build a richer dictionary of hate speech patterns that can be used, along with a unigram dictionary, to detect hateful and offensive online texts. Lack of study of the presence of hate speech among the different genders, age groups and regions, etc.

III.SYSTEM DESIGN

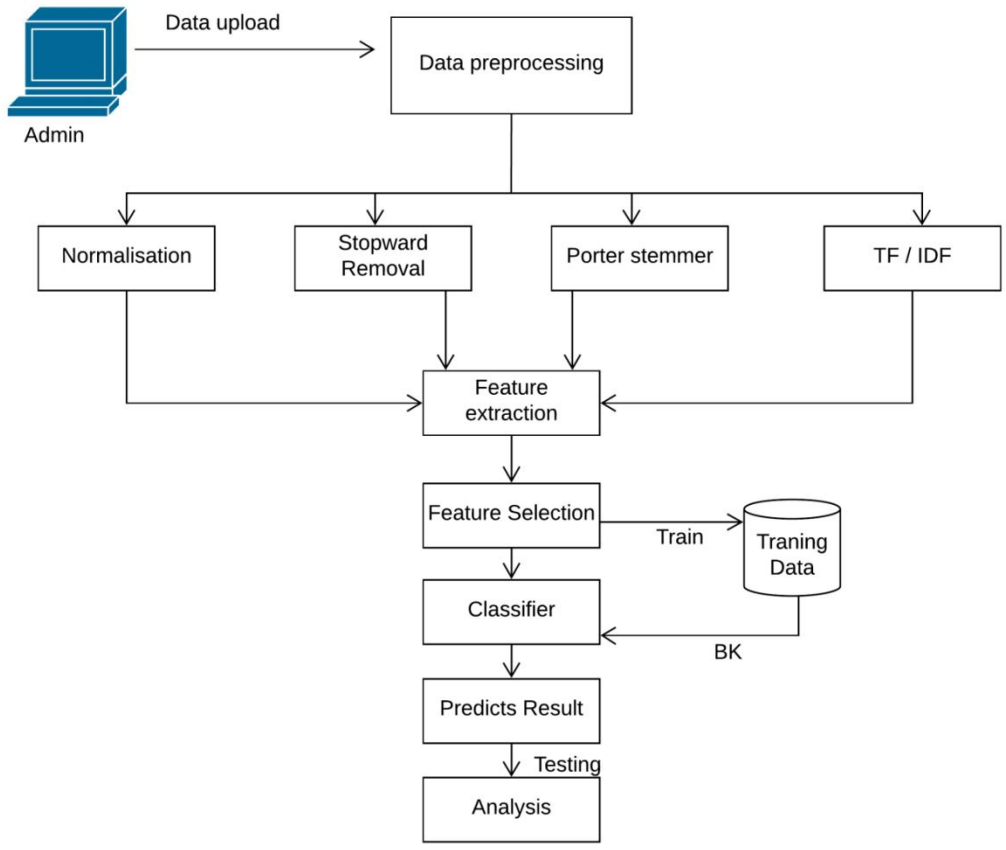

Figure 1 : Proposed System Architecture

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

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 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. Classification algorithm (SVM)
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)
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.

The below figure 2 shows the system accuracy with different machine learning algorithm helping with various Natural language Processing (NLP) features.

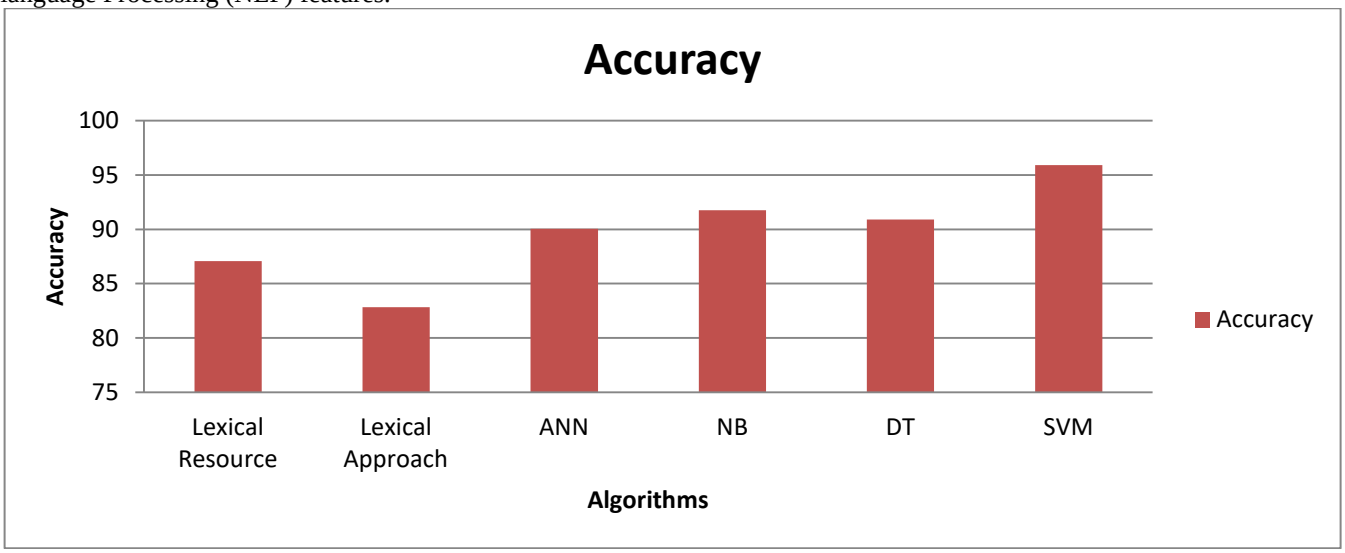

Figure 2 : System accuracy with various machine learning algorithms

Finally according the above graphs we conclude system provides better classification accuracy with SVM algorithm rather than other supervised base machine learning algorithms.

V.CONCLUSION

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. 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. 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.

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] Del Vigna12, Fabio, et al. "Hate me, hate me not: Hate speech detection on facebook." (2017).
[7] Pratiwi, Nur Indah, Indra Budi, and IkaAlfina. "Hate Speech Detection on Indonesian Instagram Comments using FastText Approach." 2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2018.APA
[8] Şahi, Havvanur, YaseminKılıç, and Rahime Belen Sağlam. "Automated Detection of Hate Speech towards Woman on Twitter."2018 3rd International Conference on Computer Science and Engineering (UBMK).IEEE, 2018.
[9] Saksesi, Arum Sucia, Muhammad Nasrun, and CasiSetianingsih."Analysis Text of Hate Speech Detection Using Recurrent Neural Network." 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC).IEEE, 2018.
[10] 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.