Machine Learning based Analysis on Human Aggressiveness and Reactions towards Uncertain Decisions

Sohaib Latif[1], Abdul Kadir Abdullahi Hasan[2], Abdaziz Omar Hassan[3]
School of Mathematics and Big Data
Anhui University of Science and Technology
Huainan, China

***Abstract*—Tweet data can be processed as a useful information. Social media sites like Twitter, Facebook, Google+ are rapidly growing popularity. These social media sites provide a platform for people to share and express their views about daily routine life, have to discuss on particular topics, have discussion with different communities, or connect with globe by posting messages. Tweets posted on twitter are expressed as opinions. These opinions can be used for different purposes such as to take public views on uncertain decisions such as Muslim ban in America, War in Syria, American Soldiers in Afghanistan etc. These decisions have direct impact on user's life such as violations & aggressiveness are common causes. For this purpose, we will collect opinions on some popular decision taken in past decade from twitter. We will divide the sentiments into two classes that is anger (hatred) and positive. We will propose a hypothesis model for such data which will be used in future. We will use Support Vector Machine (SVM), Naive Bayes (NB), and Logistic Regression (LR) classifier for text classification task. Further-more, we will also compare SVM results with NB, LR. Research will help us to predict early behaviors & reactions of people before the big consequences of such decisions.**

***Keywords*—Opinion mining; Naïve Bayes; linear regression; support vector machine**

I. Introduction

Internet is providing all the services a normal user looking for. Starting from the health, education, government and business, all categories of modern life have been covered in the shape of internet. Internet provides connectivity [1,2] between people and information publicly shared globally. Similarly, social media such as Facebook, Twitter, YouTube are platform to remain updated with current news and a airs. Through social media people [3,4,5] can share news, share their opinions and participate in activities being held online. Social Networking Sites (SNS) such as Twitter and Facebook have a beneficial effect on our way of life. SNS has been used for expressing opinions on different issues. In this work, we propose a sentiment based method for the predication of aggressive estimation.

In the age of technology [6,7] millions of people are using social media sites like Facebook, Twitter, Google Plus, etc. to share and express their views, emotions, and opinion about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media are now become rich of data in the form of tweets, status updates, posts, blog, comments, reviews, etc. [8, 9]. These social sites are not just using for personal use, but now it become a fastest tool to reach the people. It provides an opportunity for businesses by giving a platform to connect with their customers for advertising. Mostly people rely on user generated content or reviews to a great extent for decision making. The online content generated by users is too rich to analyze by normal user. The thing is to automate the process to take the views of user's as opinion. The online contents are mainly consider as opinions, sentiments, attitudes, and emotions [10].

II. Literature Review

Machine Learning, Data mining and Natural Language processing all used together for the classifications of text documents widely. These three techniques also used to discover patterns from the electronic documents. Text mining is used to discover hidden useful information from the documents and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization [11].

There have been many e orts regarding text classifications in the past. Krishna and Gonghzu [12] have analyzed large data from clinics and try to find the clinical disorders. Sonia and Shruti [15] have used Machine Learning techniques for analysis of social network E-Health data. Roshan and Rio D Souza [13] have analyzed product value using sentimental analysis publicly given on Twitter. Both have [16] worked to solve the problem of reading millions of reviews by a single user for a particular product, they have developed a model us-ng reviews posted which gives product classification in term of positive, negative and neutral reviews. In the same context, Barnaghi et al. [14] used Twitter sentiments to predict event winner. They used Bayesian Logistic Regression (BLR). They manually labelled tweets into two categories positive and negative. A model proposed [17,19] by them can be used to predict winner of any event using sentiments. In our research we will propose a methodology to analyses the pattern of human behaviors towards uncertain decisions. Our proposed methodology saves time and cost for such a huge public re-view posted daily on social networks. Nirbhay Kashyap et al. [20] have worked on music lyrics to categorize the mood of individuals. They have used different text mining and data

mining approaches to deal with such a problem. They have considered music associations, melody choice and music proposal as a feature to demonstrate the data. It is beneficial for predicting more ac-curate understanding of the music mood in the mood mapping process. Similarly, many studies have been found to investigate the online business trends using social data. Online business and larger companies' world-wide used use feedback which has been given on social sites for the improvement of product and business need with the passage of time. The amount of text and information shared on twitter in the form of tweets have valid information and it can be used to track the progress of product. They have categorized the data into different categories such as against, positive and negative and used machine learning clustering algorithms to do so. They have found that the data available online can be used for the process of information extraction and it is beneficial for the companies to track the progress of their product and handy for future considerations [21].

Santoshi et al. [22] have used twitter data differently. They have tried to figure out the user behavior towards political parties. They have captured twitter data before the election and categories the raw data into 5 different categories such as positive, negative, happy, sad and neutral. This type of information is very handy for political parties before the election. It is also effective to solve the real problems of people so that you can change the thinking of users. They have considered BJP and INC for their purpose. These are the biggest political parties in India. Using text mining and unsupervised lexical method classified tweets related to these parties to identify people emotions for the parties.

Xin Li [23] have adopted the same platform for his studies with his group mates. They have used different Natural language processing techniques [25] for the awareness of social issues human facing. Social awareness in-formation is analyzed by applying text mining and social network analysis.

AK Rathore et al. [24] has collected twitter data for the prediction of Pizza success after its launch. It is very handy information they have worked. This type of methodologies can be used to predict the behavior of any user for a particular product. Rathore and his company has used R and NodeXL for analyzing tweets collected from twitter. Furthermore, they have used different text mining, Natural Language Processing and Network Analysis techniques to predict user behavior. Any company or food delivering company can used this sort of information for t [26, 27] he purpose of success and failure of product. Nobody has worked to analyze the behavior of certain decision and their impact of human life before. In our research we will propose a methodology to analyses the pattern of human behaviors towards uncertain decisions. Our proposed methodology saves time and cost for such a huge public review posted daily on social networks.

III. Proposed Method

The solution we suggest involves Twitter data. Tweets collected with Twitter Search API [18]. Our methodology consists of two steps: training and testing phases. Feature representation and tweets collection and classifier training comes in training phase, while the testing phase have four phases: tweets collection for testing, feature representation, hypothesis prediction and evaluation. The first two tasks (i.e. tweets collection and feature representation) are shared between training and testing phase. Some popular classifiers such as SVM, NB and LR used in training and hypothesis. We have used WEKA tool for training and testing of our propose methodology. Firstly, we divided the data sets into two parts, training data and secondly testing data.

A. Preprocessing

Preprocessing reshape the data into desired form. The data collected is not purified for the process of classification, for this we have applied data Processing methodologies to transform the data into meaningful features.

Fig. 1 is showing the training of dataset. This involve mainly tokenization (or featuring), feature weighting and data cleaning (removal of irrelevant features). Once the data is collected, URLs from the tweets and replies were removed. Data only with image or with a link but there was no textual information was also re-moved. Stop words also do not give any information about topic and just create noise in the data so using stop word-list they were also removed from the data. Pre-processing is the key process in data classification tasks. It also improves the effectiveness of proposed classifier. When data is pre-processed it helps in saving classifier time while classifying. Collected tweets are further pre-processed with following steps.

1) Tokenization: Tokenization deals with breaking of long text strings into substrings which may include phrases and words collectively known as tokens. Among two ways of tokenization (phrase and word tokenization), word-level tokenization is considered as more effective due to statistical significance. In this process, the sentence for in-stance "Trump is mentally disturbed person" was bro ken into tokens "Trump", is, mentally, disturbed, person. The algorithms which are used to tokenize a sentence separate the tokens with whitespace and some are based on built in dictionary. Text can be tokenized in two ways, by words (often called bag of words) or phrases.

2) Feature Weighting: A standard function to compute the weights is TF-IDF. TF-IDF scheme is based on two parts: TF and IDF. TF stands for term frequency which is used to counts the represented terms/tokens in a document. It can give a complete measure of term occurrence. IDF stands for inverse document frequency of a term in a collection of documents.

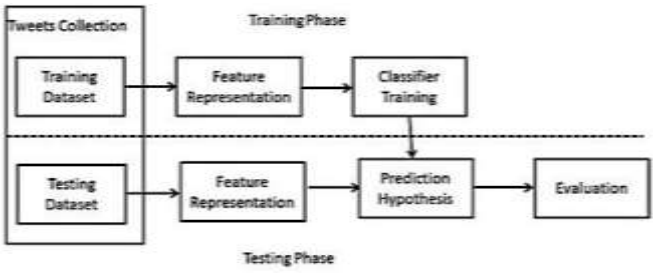

Fig. 1. Blocked Diagram of Proposed Methodology.

B. Sentiment Classification

Once we applied the pre-processing, we have data in a suitable format to apply classification algorithm on it. We have categorized the data into two formats. A data with false words labeled as Negative and data with positive words labeled as Positive. A sample of tweets rows which we have labeled. Different algorithms are available in this domain that can be used to train the classification task. Different experimental studies have been directed to analyze these methods for text categorization.

Once we applied the pre-processing, we have data in a suitable format to apply classification algorithm on it. We have categorized the data into two formats. A data with false words labeled as Negative and data with positive words labeled as Positive. A sample of tweets rows which we have labeled. Different algorithms are available in this domain that can be used to train the classification task. Different experimental studies have been directed to analyze these methods for text categorization.

IV. Classification

Supervised classification is a machine learning approach in which training data are used to construct the model and test data are used to evaluate the constructed model on unseen data to measure the performance of algorithm. There are a number of classifiers that exist to classify data, and below in Table I we will discuss the classifiers which we have explored in this work.

TABLE I. Tweets Rows with Labeled Positive and Negative Samples

Sr. No	Positive Samples	Negative Samples
1	RT @joshua_landis: Major U.S. policystatement on #Syria by Gen. Mattis: "US tofight IS in Syria until IS declares thatthey're done." Als…[War in Syria]	RT @Palespanish: Saudi Arabia: *Played a huge part in destabilizing Syria [War inSyria]
2	catalonia election spain s king felipe warnsseparatists many truly seek god s mess[Catalonia]	not only that mexican people areindigenous to north america any nativeshould be against the wall [NoBanNoWall]
3	israeli terrorist politician harasses palesoren hazan rightwing israeli known forpublicity stunts was f**k [Jerusalem]	so in your opinion jerusalem is in whichcountry [Jerusalem]
4	#Modi RafaleScam msg to cntry..thinkbig..do big.. forget ppl #NoteBan#notebandi #amitshahkiloot [Notebandi]	Bjp is likely to show sunny leone ji CD sothat people will forget about GST & NOTEBANDI [Notebandi]
5	RT @newsbusters: Says it all! The liberalmedia are STILL obsessed with trashinghttps://t.co/Y1HWx cpesP[Trump Victory]	RT @leedsgarcia: The administration let DACA renewals sit inmailboxes — and then rejected them for being "late" https://t.co/4LYHO Victory]

SVM provides better results than other Machine Learning algorithms in sense larger boundary distributions. SVM also supports high dimensional data. SVM is suitable for millions of features at the same time. SVM also supports optimization problems. Software libraries present for the implementation of SVM are lib-linear, libsvm. In logistic regression function, we have the hypothesis below, and sigmoid activation function.

Nave Bayes is probabilistic classifier which strongly based on Bayes Theorem. Simple Bayes, Independence Bayes are common names which are used. It is mostly used in classifying text information into their respective categories. There are some other example which are associated with the classifier such as to check either email is spam or not, either emails is related to sports or not.

V. Experimental Setup

A. Evaluation Measures

We used various evaluation measures to assess the results, and these measures are described below in Table II.

The results of sentiment classification using Logistic classification are given in Table III. Precision, recall, and f-measure are approximately 83%, 84%, and 84%, respectively.

Here we have given the results of sentiment classification. The results of sentiment classification using Support Vector Machine (SVM) classification are given in Table IV. Precision, recall, and f-measure are approximately 92%, 85%, and 88%, respectively.

The results of sentiment classification using Naïve Bayes (NB) classification are given in Table V. Precision, recall, and f-measure are approximately 85%, 86%, and 85%, respectively.

TABLE II. Bag of Words used for Classifications

Sr. No	Hash Tag	Words
1	#Trump	shithole, criminal trump, trump shutdown, turned sour, reject, failed socialist, evils choice, Moscow's victory, what nonsense, shocking crimes, disgrace, slap, Pathetic
2	#SaudiWomenCanDrive	condemned, cheesy, car accident, lack of intellect, Protests ,condemnation, break,license, disasters
3	#PanamaVerdict	squeezed, shameless, Patwari, Haram Family, corruption, trashed, bark, barking, gangster, anti, wolf
4	#NoteBandi	waste of time, economy mess up, destroyed, black money, laundering, su er, cor-ruption, e ect, common man, impact, a ected, self-destructive, slap, idiot, com-plaining, ruthless, corrupted gov, history, disasters, stunts, nashbandi, Nobandi
5	#NoBanNoWall	attacks, hurt, worse, wasteful, monument, hurdle, damage, environment, hate, break the wall, harmed, wreck less, darkness, must resist, rapists, hurt brownpeople, discriminatory ban stupid wall, resist Trump

TABLE III. RESULTS OBTAINED USING LOGISTIC REGRESSION

Sr. No	Class	Precision	Recall	F-Measures
1	Brexit	0.814	0.763	0.784
2	Catalonia	0.879	0.885	0.882
3	PanamaVerdict	0.881	0.872	0.876
4	NoteBandi	0.86	0.822	0.844
5	Jerusalem	0.86	0.817	0.834
6	SaudiWomenCanDrive	0.857	0.85	0.855
7	Trump	0.856	0.864	0.86
8	MuslimBan	0.855	0.854	0.856
9	NoBanNoWall	0.89	0.891	0.891
10	SyriaWar	0.863	0.879	0.869

TABLE IV. RESULTS OBTAINED USING SVM CLASSIFIER

Sr. No	Class	Precision	Recall	F-Measures
1	Brexit	0.865	0.883	0.863
2	Catalonia	0.841	0.908	0.873
3	PanamaVerdict	0.922	0.925	0.906
4	NoteBandi	0.900	0.899	0.889
5	Jerusalem	0.926	0.929	0.921
6	SaudiWomenCanDrive	0.911	0.900	0.901
7	Trump	0.860	0.817	0.834
8	MuslimBan	0.850	0.873	0.843
9	NoBanNoWall	0.916	0.924	0.909
10	SyriaWar	0.922	0.921	0.905

TABLE V. RESULTS OBTAINED USING NAÏVE BAYES

Sr. No	Class	Precision	Recall	F-Measures
1	Brexit	0.817	0.821	0.819
2	Catalonia	0.864	0.846	0.855
3	PanamaVerdict	0.866	0.856	0.862
4	NoteBandi	0.899	0.889	0.879
5	Jerusalem	0.872	0.886	0.875
6	SaudiWomenCanDrive	0.866	0.867	0.865
7	Trump	0.819	0.815	0.817
8	MuslimBan	0.798	0.824	0.809
9	NoBanNoWall	0.875	0.880	0.878
10	SyriaWar	0.864	0.885	0.869

Precision (Positive Predictive value) can be defined as relevant instances from the retrieved instances. The concept is used for binary classifications. Whereas recall is the number of relevant instances from total number of relevancy. This is also known as sensitivity.

To get good performance of classifier precision and recall are often used together [28]. F-Measure can be defined as harmonic mean of precision and recall.

B. Tools for Evaluation

To perform desired task, we used WEKA. WEKA is open source free software which has been used for various machine learning problems using data. It contains tools which can be used for classifications, pre-processing, clustering, visualization, association rules etc. Machine Learning is nothing without giving an artificial intelligence to your data. Machine learning methods are very similar to data mining algorithms. WEKA has collection of Machine Learning (ML) algorithms which are applied on data to extract desired results from it.

C. Comparative Analysis

A comparison analysis of classifiers for sentiment classification is given in Table VI. We can see that SVM provides best results and it gives approximately 88% F-measure which is much better than from NB and LR results.

TABLE VI. COMPARATIVE ANALYSIS OF RESULTS OBTAINED USING ALL THREE CLASSIFIERS

Class	Precision	Recall	F-Measures
SVM	0.92	0.901	0.899
NB	0.807	0.811	0.801
LR	0.799	0.789	0.799

VI. DISCUSSIONS

In last chapter we have described tools, data source, and different technologies that we have used in our approach. In this chapter we will present the obtained results. Fig. 2 is showing the work flow of our experimentation. Three classifiers Support Vector Machine, Naive Bayes and Logistic Regression are used in our experiment and to measure the e effectiveness of each classifier we have used three measurements i.e. recall, precision, and f-measure by applying standard 10-folded cross-validation.

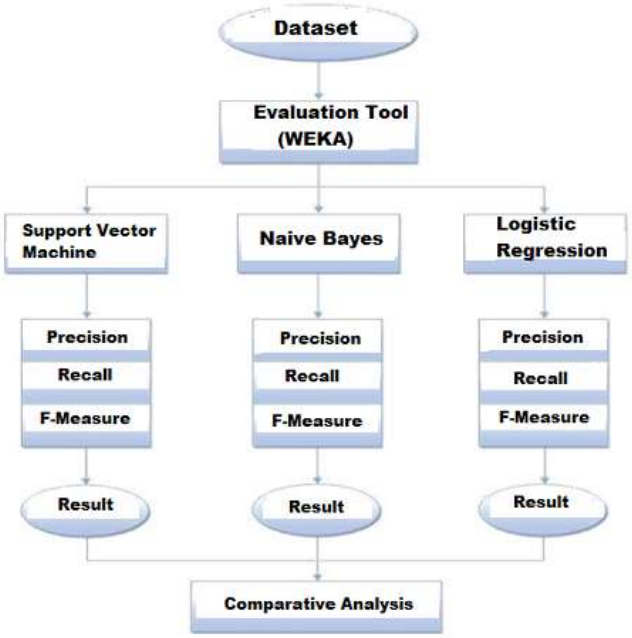

Fig. 2. Workflow of Experimentation.

VII. CONCLUSION

Twitter is one of the most important social sharing platform for useful information. Tweets posted on twitter are expressed as opinions. These opinions can be used for different purposes such as to take public views on uncertain decisions such as Muslim ban in America, War in Syria, American Soldiers in Afghanistan, etc. These decisions have direct impact in users life such as violations & aggressiveness are common causes. We have collected tweets of such decisions and labeled the tweets into two categories such as anger (hatred) and positive. We have used classifier algorithms such as Sup-port Vector Machine (SVM), Naive

Bayes (NB), and Logistic Regression (LR) for building models. We have also compared SVM results with NB, LR. This research is useful for predicting early behaviors & reactions of people before the big consequences of such decisions.

In the future we interested to build a tool which can work as a recommender system to classify tweets automatically into two categories such as Anger and Positive.

$a^2 + b^2 = c^2$

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China (No.61572035, No.61902002, No.61402011 and No.61272153), the Natural Science Foundation of Educational Government of Anhui Province of China(No.KJ2016A208), Anhui Provincial Big Data Foundation(No. 2017032), the Foundation for top-notch academic disciplines of Anhui Province(No.gxbjZD11),the Academic and Technology Leader Foundation of Anhui Province(No.2019H239), the Open Project Program of Key Laboratory of Embedded System and Service Computing of Ministry of Education(No.ESSCKF2018-04). We also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentations.

In the end, researcher would like to pay thanks to my father and mother, especially because both of them suffer a lot during my studies.

REFERENCES

[1] Liu.B, Sentiment analysis and opinion mining Synthesis lectures on human language technologies, 5(1), 1-167, 2012.

[2] Khan.F.H, Bashir.S. And Qamar.U. TOM: Twitter opinion mining framework using hybrid classification scheme Decision Support Systems, 57, 245-257, 2014.

[3] Lin.C, And He.Y. Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 375-384, ACM, 2009.

[4] Jiang. L, Yu. M, Zhou. M, Liu.X and Zhao.T .Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- Association for Computational Linguistics .Volume 1, pp. 151-160, 2011.

[5] Barbosa.L, and Feng. J. Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics Association for Computational Linguistics Posters, pp. 36-44. 2010.

[6] Torunoglu.D, Telseren.G, Sagturk.O and Ganiz.M.C. Wikipedia based semantic smoothing for twitter sentiment classification. In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on pp. 1-5, IEEE. 2013.

[7] Go. A, Bhayani.R and Huang.L. Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford, 1(2009), 12. 2009.

[8] Liu. K.L, Li.W.J and Guo.M. Emoticon smoothed language models for twitter sentiment analysis In AAAi.publications, 2012.

[9] Bravo Marquez.F, Mendoza.M and Poblete.B Combining strengths, emotions and polarities for boosting Twitter sentiment analysis. In Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (p. 2) ACM, 2013.

[10] Goncalves.P, Araujo.M, Benevenuto.F and Cha.M. Comparing and combining sentiment analysis methods. In Proceedings of the first ACM conference on Online social networks pp. 27-38, ACM, 2013, October.

[11] Khan. A, Baharudin.B, Lee.L. H and Khan, K. A review of machine learning algorithms for text-documents classification. Journal of advances in information technology 1(1), 4-20, 2010.

[12] Chodey.K. P and Hu.G. Clinical text analysis using machine learning methods. In Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on pp. 1-6. IEEE. 2016.

[13] Fernandes.R and D'Souza. R Analysis of product Twitter data though opinion mining. In India Conference (INDICON), 2016 IEEE Annual, pp. 1-5 .IEEE, 2016.

[14] Barnaghi.P, Ghaffari.P and Breslin.J. G. Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. In Big Data Computing Service and Applications (Big Data Service), 2016 IEEE Second International Conference on pp. 52-57, IEEE, 2016.

[15] Saini.S and Kohli.S. Machine learning techniques for effective text analysis of social network E-health data. In Computing for Sustainable Global Development (INDIA.Com), 2016 3rd International Conference on pp. 3783-3788, IEEE, 2016.

[16] Nawaz.M.S, Bilal.M, Lali.M.I. Ul Mustafa.R, Aslam.W and Jajja.S. Effectiveness of Social Media Data in Healthcare Communication. Journal of Medical Imaging and Health Informatics, 7(6), 1365-1371, 2017.

[17] Liu.B and Zhang.LA survey of opinion mining and sentiment analysis. In mining text data Springer US pp. 415-463, 2012.

[18] Twitter Search API. Retrieved from website: https://dev.Twitter.com/rest/public/search, 2015.

[19] Hall.M, Frank.E, Holmes.G, Pfahringer.B, Reutemann.P and Witten. I. H. The WEKA data mining software an update ACM SIGKDD explorations newsletter, 11(1), 10-18 ,2009.

[20] Kashyap, Nirbhay et al. Mood Based Classification of Music by Analyzing Lyrical Data Using Text Mining Micro-Electronics and Telecommunication Engineering (ICMETE), 2016 International Conference on. IEEE, 2016.

[21] Zunic, Emir, AlmirDjedovic, and DzenanaDonko. Application of Big Data and text mining methods and technologies in modern business analyzing social networks data about traffic tracking. Telecommunications (BIHTEL), 2016 XI International Symposium on IEEE, 2016.

[22] Kuamri, Santoshi and NarendraBabu.C. Real time analysis of social media data to understand people emotions towards national parties, 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT).IEEE, 2017.

[23] Vespignani, Alessandro. Predicting the behavior of techno-social systems. Science 325.5939, 425-428, 2009.

[24] Rathore, Ashish Kumar and VigneswaraIlavarasan.P. Social media analytics for new product development Case of a pizza. Advances in Mechanical, Industrial Automation and Management Systems (AMIAMS), 2017 International Conference on IEEE, 2017.

[25] Ricardo. A, Calix, Automated semantic understanding of human emotions in writing and speech, 2011.

[26] Pang.B , Lee.L, Vaithyanathan.S, Thumbs up sentiment classification using machine learning techniques .in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 9-86, 2002.

[27] Janyce.M.Wiebe, Bruce.Rebecca.F, O'Hara and Thomas P. Development and use of a gold-standard data set for subjectivity classifications in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 246—253, 1999.

[28] Full list of bad words and top swear banned by Google Available at: https://www.freewebheaders.com/full-list-of-bad-words-banned-by-google/Lastaccessed ,2019.