2020 International Conference on Computational Science and Computational Intelligence (CSCI)

Cyberbullying Detection Through Sentiment Analysis

Jalal Omer Atoum
Department of Mathematics and Computer Science,
East Central University
Ada, Oklahoma
jomer@ecok.edu

Abstract-In recent years with the widespread of social media platforms across the globe especially among young people, cyberbullying and aggression have become a serious and annoying problem that communities must deal with. Such platforms provide various ways for bullies to attack and threaten others in their communities. Various techniques and methodologies have been used or proposed to combat cyberbullying through early detection and alerts to discover and/or protect victims from such attacks. Machine learning (ML) techniques have been widely used to detect some language patterns that are exploited by bullies to attack their victims. Also. Sentiment Analysis (SA) of social media content has become one of the growing areas of research in machine learning. SA provides the ability to detect cyberbullying in real-time. SA provides the ability to detect cyberbullying in real-time. This paper proposes a SA model for identifying cyberbullying texts in Twitter social media. Support Vector Machines (SVM) and Naïve Bayes (NB) are used in this model as supervised machine learning classification tools. The results of the experiments conducted on this model showed encouraging outcomes when a higher n-grams language model is applied on such texts in comparison with similar previous research. Also, the results showed that SVM classifiers have better performance measures than NB classifiers on such tweets.

Keywords — Cyberbullying, sentiment analysis, machine learning, social media

I. INTRODUCTION

Social media has been used by almost all people especially young adults as a major media of communication. In [1], young adults were among the earliest social media adopters and continue to use it at high levels, also, usage by older adults has increased in recent years as shown in Fig. 1.

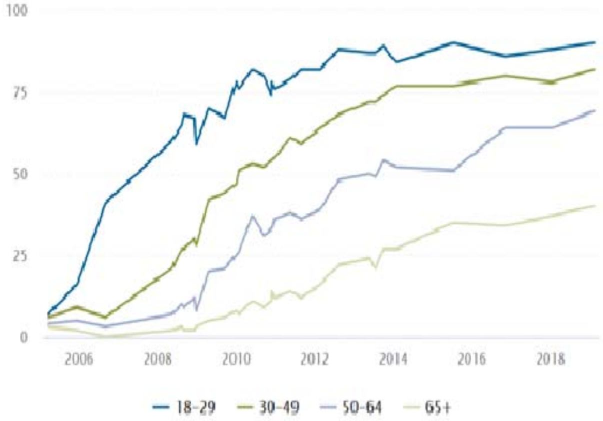

Figure 1. Percentage of U.S. Adults Who Use Social Media Sites by Age.

As a result of such wide usage of social media among adults, cyberbullying or cyber aggression has become a major problem for social media users. This had lead to an increasing number of cyber victims who have suffered either physically, emotionally, mentally, and/or physically.

Cyberbullying can be defined as a type of harassment that takes place online on social networks. Criminals rely on such networks to collect data and information to enable them to execute their crimes, for example, by determining a vulnerable victim [2]. Therefore, researchers have been working on finding some methods and techniques that would detect and prevent cyberbullying. Recently, monitoring systems of cyberbullying have gained a considerable amount of research, their goal is to efficiently identify cyberbullying cases [3]. The major idea behind such systems is the extraction of some features from social media texts then building classifier algorithms to detected cyberbullying based on such extracted features. Such features could be based on users, content, emotions, and/or social networks. Furthermore, machine learning methods have been used to detect language pattern features from texts written by bullies.

The research in detecting cyberbullying has been mostly done either through filtration techniques or through machine learning techniques. Infiltration techniques, profane words or idioms have to be detected from texts to identify cyberbullying [4]. Filteration techniques usually use Machine learning methods to build classifiers that have the capabilities of detecting cyberbullying using corpora of collected data from social networks such as Facebook and Twitter. For instance, in [5], data were collected from Formspring then it was labeled using the Amazon Mechanical TURK [6]. WEKA toolkit [7] machine learning methods were, also, employed to train and test these classifiers. Such techniques suffer from an inability to detect indirect language harassment [8].

Chen [9] had proposed a technique to detect offensive language constructs from social networks through the analysis of features that are related to the users writing styles, structures, and certain cyberbullying contents to identify potential bullies. The basic technique used in this study is a lexical syntactic feature that was successfully able to detect offensive contents from texts sent by bullies. Their results had indicated a very high precision rate (98.24%), and recall of 94.34%.

Nandhini and Sheeba [10] had proposed a technique for detecting cyberbullying based on an NB classifier using data collected from MySpace. They had reported an achieved accuracy of 91%. Romsaiyud el a. [11] had employed an enhanced NB classifier to extract cyberbullying words and clustered loaded patterns. They had achieved an accuracy of

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00056

95.79% using a corpus from Slashdot, Kongregate, and MySpace.

In this research, we use the Sentiments Analysis (SA) method for the classification of tweets into either positive, negative, or neutral concerning cyberbullying. We proposed a technique for preprocessing tweets, then we tested and trained two supervised machine learning classifiers, namely; Support Vector Machine (SVM) and Naïve Bayes (NB). Then we compared our proposed technique with other similar work presented in [12].

Section two presents the background for this research. Section three presents the proposed tweets sentiment analysis model. Section four presents the experiments and results of this proposed model. Finally, section five presents the conclusions of this research.

II. BACKGROUND

Machine learning (ML) is a method of data analysis that automates analytical model building. ML algorithms are often categorized as supervised or unsupervised. Supervised ML algorithms apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. Unsupervised ML algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The problem with unsupervised ML is that they may overlap and learn to localize texts with minimal unsupervised algorithms. Many researchers have used supervised learning approaches on data related to publicly released corpora [13].

Naïve Bayes (NB) classifiers as supervised learning models are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models. NB often relies on the bag of words presentation of a document, where it collects the most used words neglecting other infrequent words. The bag of words depends on the feature extraction method to provide the classification of some data [14]. Furthermore, NB has a language modeling that divides each text as a representation of unigram, bigram, or n-gram and tests the probability of the query corresponding with a specific document.

Support-Vector Machines (SVMs) are also supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, a SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting) [15]. The most important models for SVM text classifications are Linear and Radial Basis functions. Linear classification tends to train the dataset then builds a model that assigns classes or categories [16]. It represents the features as points in space predicted to one of the assigned classes. SVM has good classification performance in several fields, but mostly applied for image recognition and text classification.

Classifiers for SA are usually based on predicted classes and polarity, and/or on the level of classification (sentence or document). Lexicon based SA text extraction is annotated with semantic orientation polarity and strength. SA proved that light stemming comes in handy for the accuracy and for the performance of classification [17].

An automatic classifier of text documents-based NB and SVM algorithms was presented in [18], the results indicated that the SVM algorithm handled the text documents classification better than the NB algorithm. Therefore, in this research and for the SA proposed techniques, we have used the two supervised ML approaches for the classifications of social media texts, namely, Naïve Bayes (NB) and Support Vector Machine (SVM).

To evaluate our classifiers, several evaluation metrics could be used. We have adopted the most common criteria that are commonly used, namely; accuracy, precision, recall, F-measure, and Receiver Operating Characteristics (ROC). Such criteria are defined as follows:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F-measure = 2 * (Recall * Precision) / (Recall + Precision)

ROC: is a plot of the TP rate against the FP rate

Where:

TP (True Positive) is a hit; correctly classified as positive.

TN (True Negative) (TN) is a rejection; correctly classified as negative.

FP (False Positive) is a false alarm, falsely classified as positive.

FN (False Negative) is a miss, falsely classified as negative.

III. PROPOSED TWEETS SA MODEL

The proposed SA model analyzes, mines, and classifies tweets. Several preprocessing stages must be done on the collected tweets for the SA process to be more effective as illustrated in Fig. 2. These stages are as follows:

A. Collecting Tweets:

A connection to Twitter is created to collect a corpus of tweets. A read-only application is built to collect written tweets from Twitter. Tweets extraction helps in extracting the important content of a tweet (the essence). Hence, what is needed from a tweet is written after the hashtags, and subsequently extracting the feature words, words that carry a message for the user whether it is a positive, negative, or neutral cyberbullying tweet. Also, tweets extraction is needed to facilitate analyzing the features vector and selection process (unigrams, bigrams, trigrams, …, n-gram), and to facilitate the classification of both training and testing sets of tweets.

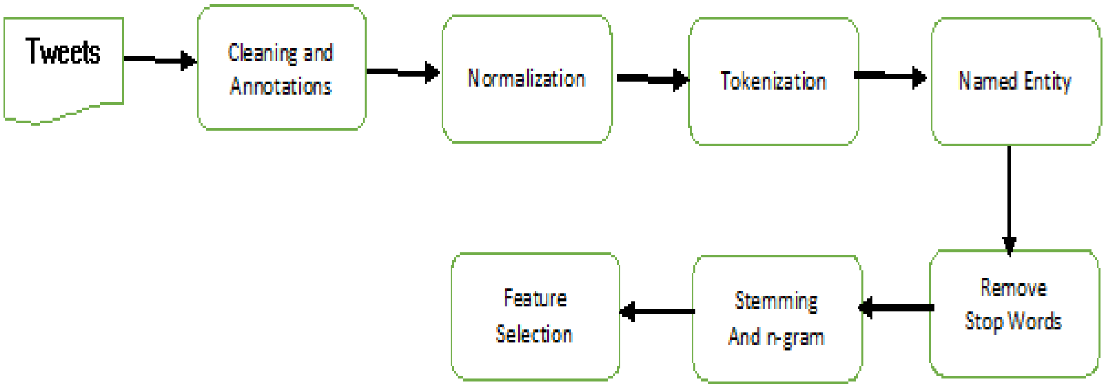

Figure 2. Proposed Tweets Preprocessing Stages**.**

B. Cleaning and Annotations of Tweets:

Cleaning the tweets by removing special symbols, and various characters and emoticons. Those symbols and characters may lead us to a different classification from what the user is intended originally in the tweet. Hence, we replace special symbols, emotions, and emotional characters with their meanings. Table 1 presents some special symbols that we have used along with their meanings and sentiments.

Furthermore, all “http/https‟ shortening, and special symbols such as (*, &, $, %, -, _,><) are removed from the collected tweets. Then each special character is replaced with a space character.

The annotation process of the collected tweets was done manually. As a result of this annotation, each tweet is labeled with either positive, negative, or neutral cyberbullying. Finally, the cleaned annotated extracted tweets are stored in a database in a comma-separated values format for further manipulation.

TABLE 1. Sample Special Symbols and Their Meanings.

Character/Symbol	Meaning	Sentiment
♥	Heart or love	Positive
☺	Smile	Positive
☹	Sad	Negative
❄	Snow	Positive or negative
✈	Bird or Airplane	Neutral
?	Question	Neutral

C. Normalization:

The normalization stage starts by removing all extra spaces. All non-standard words that have numbers and/or dates are identified. Such words would be mapped into especially built-in vocabularies. This results in a smaller number of tweet vocabularies and improves the accuracy of the classification task.

D. Tokenization:

Tokenization is an important step in SA since it reduces the typographical variation of words. The feature extraction process and the bag of words require tokenization. A dictionary of features is used to transform words into feature vectors, or feature indices; such that the index of the feature (word) in the vocabulary is linked to its frequency in the whole training corpus.

E. Named Entity Recognition (NER):

NER is a significant tool in natural language processing; it allows the identification of proper nouns in an unstructured text. NER has three categories of name entities; ENAMEX (person, organization, and country), TIMEX (date and time), and NUMEX (percentages and numbers).

F. Removing stop words:

Some stop words can help in attaining the full meaning of a tweet and some of them are just extra characters that need to be removed. Some examples of stop words are: "a," "and," "but," "how," "or," and "what.", such stop words do not affect the tweets meaning and can be removed from tweets.

G. Stemming:

Tweets stemming is done by removing any attached suffixes, prefixes, and/or infixes from words in tweets. A stemmed word represents a broader concept of the original word, also it may lead to save storage [19]**.** The goal of stemming tweets is to reduce the derived or inflected words into their stems, base, or root form in order to improve SA. Furthermore, stemming helps in putting all the variation of a word into one bucket, effectively decreasing our entropy and gives better concepts to the data. Moreover, N-gram is a traditional method that takes into consideration the occurrences of N-words in a tweet and could identify formal expressions [20]. Hence, we have used N-gram in our SA.

In this research, we have implemented the term frequency using *weka* [21]. Term frequency assigns weights for each term in a document in which it depends on the number of occurrences of the term in a document, and it gives more weight to those terms that appear more frequent in tweets because these terms represent words and language patterns that are more used by the tweeters.

H. Feature Selection

Feature selection techniques have been used successfully in SAs [22] [23]. In which Features would be ranked according to some measures such that non useful or non-informative features would be removed to improve the accuracy and efficiency of the classification process. In this study, we have used the Chi-square and Information gain techniques to remove such irrelevant features.

IV. EXPERIMENTS AND RESULTS

To evaluate the performance of the machine learning methods used in this research; namely the Naïve Bayes (NB) and the Support Vector Machine (SVM), we have collected a total of 5628 tweets (Positive-cyberbullying, negative-no cyberbullying, and neutral). This set of tweets was manually classified into 1187 cyberbullying tweets, 2342 with no cyberbullying tweets and the remaining 2099 are neutral tweets. Table 2 presents the distribution of these tweets.

Before conducting our experiments, the set of tweets had gone through the various phases of cleaning, preprocessing, Normalization Tokenization, Named Entity Recognition, stemming, and features selection as has been discussed in the previous section. Then this data set is split into a ratio of (70, 30) for training and testing the NB and SVM classifiers. Finally, cross-validation is used in which 10-fold equal-sized sets are produced.

Several experiments have been conducted to compare the performance of NB and SVM classifiers of the above-collected set of tweets. In the first experiment, tweets with 2-gram, 3-gram, and 4-gram are used to evaluate the NB and SVM classifiers in terms of accuracy, precision, recall, F-measure, and ROC. Table 3 presents the results of this experiment. Fig. 3 illustrates the averages of the measures obtained over the different n-grams models for both NB and SVM classifiers. From Table 3 and Fig. 3 we can conclude that SVM classifiers have achieved higher average results than the NB classifiers in all n-gram language models in terms of accuracy, precision, recall, F-measure, and ROC. For instance, SVM classifiers achieved an average accuracy value of 92.02% in the case of the 4-gram language model, whereas, the NB classifiers achieved an average accuracy of 81.1 on the same language model. Also, the 4-gram language model has outperformed all other n-grams language models in all measures in both SVM and NB classifiers. This is because a higher n-gram leads to an increase in the probability of estimation.

TABLE 2. Tweets Statistics

Total number of Tweets	5628
Number of positive (cyberbullying) Tweets	1187
Number of negative (no cyberbullying) Tweets	2342
Number of neutral Tweets	2099

TABLE 3. NB and SVM Measures for Different N-gram Language Models

Measure		2 gram	3 gram	4 gram	Average
Accuracy	NB	82.35	81.7	81.1	82.025
	SVM	91.21	91.7	92.02	91.64
Precision	NB	78.46	78.68	78.42	78.52
	SVM	88.92	89.1	89.3	89.11
Recall	NB	77.31	79.4	79.71	78.81
	SVM	86.28	87.36	88.04	87.23
F-Measure	NB	77.88	79.04	79.06	78.66
	SVM	87.58	88.22	88.66	88.16
ROC	NB	78.61	77.9	78.03	77.9
	SVM	88.2	88.56	89.3	88.93

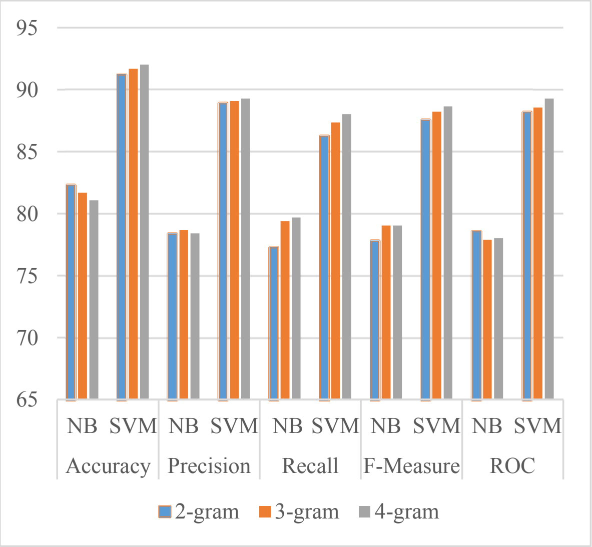

Figure 3. Graphical Comparisons of NB and SVM Measures

Another experiment was conducted to compare our proposed classifiers to the work presented in [12] using the two major classification techniques, namely; Naïve Byes (NB), and Support Vector Machine (SNM) using the same data set presented earlier. Table 4 presents the summarized performance measures of our proposed techniques in implementing the NB and SVM classifiers in comparison with the implementation of [12]. It is very clear from Table 4 and Fig. 4, that in most measures we had obtained slightly better results.

TABLE 4. Averages of NB and SVM Measures for Different N-gram Language Models

		Avg. Accur.	Avg. Recall	Avg. Prec.	Avg. F-Meas.	Avg. ROC
Proposed Work	NB	81.71	78.8	78.52	78.65975	77.9
	SVM	91.64	87.22	89.1	88.14997	88.93
Previous work	NB	80.9	79.1	77.04	78.05641	77.02
	SVM	83.46	85.3	84.32	84.80716	85.71

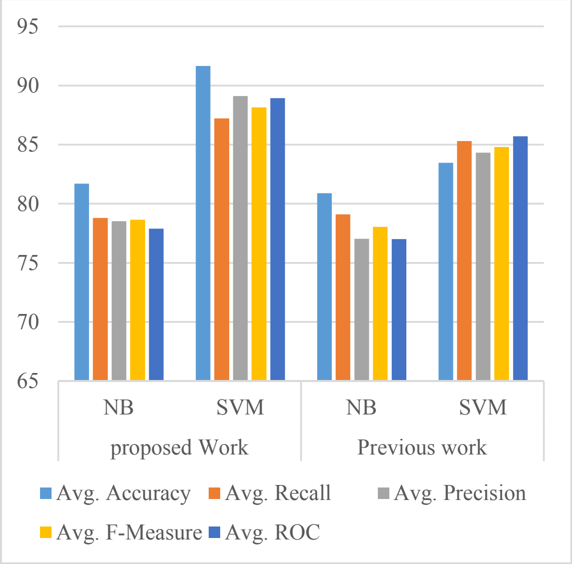

Figure 4. Averages of Graphical Comparisons of NB and SVM Measures

Furthermore, as shown in Fig. 4, the performance measures of the our SVM classifiers have better results than the SVM classifiers of the previous work. For instance, we have obtained an average accuracy of 91.61 in the proposed work in contrast of an average accuracy average of 83.44 in the previous work. Also, the average ROC of our SVM classifier is 88.93 compared to 85.71 of the SVM of the previous work.

This is an impressive result since ROC compares the true positive and false-positive rates, which is the fraction of the sensitivity or recall in machine learning.

V. CONCLUSIONS

In this research, we have proposed an approach to detect cyberbullying from Twitter social media platform based on Sentiment Analysis that employed machine learning techniques; namely, Naïve Bayes and Support Vector Machine. The data sets used in this research is a collection of tweets that have been classified into positive, negative, or neutral cyberbullying. Before training and testing such machine learning techniques, the collected set of tweets have gone through several phases of cleaning, annotations, normalization, tokenization, named entity recognition, removing stopped words, stemming and n-gram, and features selection.

The results of the conducted experiments have indicated that SVM classifiers have outperformed NB classifiers in almost all performance measures over all language models. Specifically, SVM classifiers have achieved an average accuracy value of 92.02%, while, the NB classifiers have achieved an average accuracy of 81.1 on the 4-gram language model.

Furthermore, more experiments have been conducted to evaluate our proposed work to a similar work of [12]. These experiments had also indicated that our SVM and NB classifiers had slightly better performance measures when compared to this previous work.

Finally, for direction research in cyberbullying detection, we would like to explore other machine learning techniques such as Neural Networks and deep learning, with larger sets of tweets. Also, to adopt some proven methods for an automated annotation process to handle such a large set of tweets.

ACKNOWLEDGMENT

I would like to gratefully acknowledge the support of East Central University-Ada, Oklahoma and the Department of Mathematics and Computer Science at ECU for their support in providing the funds for my paper registration in this conference.

REFERENCES

[1] JUNE 12, 2019, PEW Research center, Internet & Technology-Social Media Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/, accessed March 28, 2020.

[2] Tavani, Herman. T., "Introduction to Cybernetics: Concepts, Perspectives, and Methodological Frameworks", In H. T. Tavani, ethics and Technology: Controversaries, questions, and Strategies for ethical Computing, river University – Fourth Edition, Wiley, pp 1-2, 2013.

[3] S. Salawau, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A survey," Vol. 3045, no c, pp 1-20, 2017.

[4] Internet Monitoring and Web Filtering Solutions", "PEARL SOFTWARE, 2015. Online. Avaliable:http://www.pearlsoftware.com/solutions/cyber-bullying-inschools.html. [Accessed Feb 20, 2020]

[5] K. Reynolds, "Using Machine Learning to Detect Cyberbullying", 2012.

[6] Amaon Mechanical Turk", Aug. 15, 2014 [Online]Available: http://ocs.aws.amazon.com/AWSMMechTurk/latest/AWSMechanical-TurkGetingStartedGuide/SvcIntro.html. Accessed July 3,2020.

[7] S. Garner, Weka: The Waikato Environment for Knowledge Analysis", New Zealand, 1995.

[8] V. Nahar, X. Li and C. Pang, "An effective Approach for Cyberbullying Detection," in Communication in Information Science and Management Engineering, May 2013.

[9] Chen, Y., Zhou, Y., Zhu, s. and Xu, H., "Detecting Offensive Language in Social Media to Protect Adolescent Online Saftey", In privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (SocialCom), pp 71-80, 2012.

[10] B. Sri Nandhinia, and J.I. Sheeba, "Online Social Network Bullying Detection Using Intelligence Techniques", International Conference on Advanced Computing Technologies and Applications (ICACTA-2015), Procedia Computer Science 45 (2015) 485 – 492

[11] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasertslip, Piyapon Nurarak, and Pirom konglerd," Automated Cyberbullying Detection Using Clustering Appearance Patterns", in Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242-247, IEEE, 2017.

[12] Dipika Jiandani, Riddhi Karkera, Megha Manglani, Mohit Ahuja, Mrs. Abha Tewari, "Comparative Analysis of Different Machine Learning Algorithms to Detect Cyber-bullying on Facebook", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 6 Issue IV, April 2018, pp. 2322-2328.

[13] Cristina Bosco and Viviana Patti and Andrea Bolioli, "Developing Corpora for Sentiment Analysis: The Case of Irony and Senti–TUT, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), pp. 4158-4162.

[14] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers", Machine learning, Vol. 29, No. 2–3, pp. 131-163, 1997.

[15] Cortes, Corinna; Vapnik, Vladimir N., "Support-Vector Networks" (PDF). Machine Learning. 20 (3): 273–297. (1995), Cutesier 10.1.1.15.9362. doi:10.1007/BF00994018.

[16] C. Cortes, and V. Vapnik. "Support-vector networks". Machine Learning, Vol. 20, No. 3, pp. 273–297,1995 doi:10.1007/BF00994018.

[17] Leimin Tian, Catherine Lai, and Johanna D. Moore, "Polarity and Intensity: The Two Aspects of Sentiment Analysis", Proceedings of the First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), pages 40–47, Melbourne, Australia July 20, 2018. 2018, Association for Computational Linguistics.

[18] Monali Bordolo, and Saroj Kr. Biswas, "Sentiment Analysis of Product using Machine Learning Technique: A Comparison among NB, SVM and MaxEnt", July 2018, International Journal of Pure and Applied Mathematics 118(18):71-83

[19] Brajendra Singh Rajput, and Nilay Khare, "A Survey of Stemming Algorithms for Information Retrieval", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 17, Issue 3, Ver. VI (May – Jun. 2015), PP 76-8

[20] L. Chen, W. Wang, M. Nagaraja, S. Wang, and A. Sheth, "Beyond Positive/Negative Classification: Automatic Extraction of Sentiment Clues from Microblogs,", Kno.e.sis Center, Technical Report, 2011.

[21] G. Holmes, A. Donkin, and I. Witten, "WEKA: A Machine Learning Workbench," In Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, Brisbane, 29 November-2 December 1994, 357-361.

[22] Khalifa, K., and Omar, N., "A Hybrid Method Using Lexicon-Based Approach and Naïve Bayes Classifier for Arabic Opinion Question Answering," Journal of Computer Science 10 (11): 1961-1968, 2014, ISSN: 1549-3636.

[23] Fattah MA, "A Novel Statistical Feature Selection Approach for Text Categorization. J Inf Process Syst 13:1397–1409. (2017), https://doi.org/10.3745/JIPS.02.0076

[24] Guyon I, Elisseeff A, "An Introduction to Variable and Feature Selection". J Mach Learn Res 3:1157–1182. (2003) https://doi.org/10.1016/j.aca.2011.07.027