Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
CASE STUDY
Available Online at www.ijarcs.info

Toxic Comment Tools: A Case Study

Pooja Parekh
Assistant Professor
Smt. Chandaben Mohanbhai Patel Institute of Comp App,
Charotar University of Science and Technology,
Changa, Gujarat, India

Hetal Patel
Assistant Professor
Smt. ChandabenMohanbhai Patel Institute of Comp App,
Charotar University of Science and Technology,
Changa, Gujarat, India

Abstract:The online web tools enable everyone who can confidently voice their opinions in public sphere. The opinion may be encouragement, blessing or good suggestion i.e. positive, or it may be negative to that extent that it must be restricted at some point. The aim of this paper is to survey the different machine learning techniques employed within the scope of discovering the hateful language on social networking site, their challenges to provide a solution to detect the toxic comment and modify it of the same.

Keywords:Toxic Comments, Offensive, Machine Learning.

I. INTRODUCTION

As the popularity of the interactive media increasing tremendously, the noticing intolerable Comment on social networking site becomes a far-reaching and vital research area. Social Networking Site where plenty of people can do discussion globally with anyone in the world. The discussion normally done in the form of comments, feedback, review and other form, which may be positive or negative. The positive text does not give any wrong impact but the major challenging problem is the negative text, and it is termed as toxic. The toxic is defined as "an awkward insolent, impolite, or comment that is likely to make one leave a discussion" [1].

The 'Tweens, Teens and Technology 2014 Report' by McAfee [2] suggests that almost half of the total Indian Youth using social media experienced toxic comments. In addition, a survey [2] reveals that the major reason behind significant number of suicides committed by teens are the negative Comments to them in social media. These comments result into increased frustration to them, which restricts them from interacting with peers and partying. Apart from this, it often results in emotional trauma for the prey. On networking sites like Myspace, mostly teenager have reveal the exasperation as it creates too much emotional stress [3]. Individuals are given compete free hand as to what they can post when online. They are also given the power to post offensive remarks or pictures regardless of what disastrous implications it may create. With the improvement in speed quality of internet usage from 2G to 5G, now in fraction of time, information can be spread around the world [4].

These Interactive media also fails in filtering these adverse comments and status which can be uploaded in public platforms, but they are equipped with reporting systems, which enables the user to report the contents as abuse, and such contents can then be removed from social platform [4].

For examples, Facebook has number of employees working on the contents, which are being uploaded daily on user's walls,profiles, comments etc. They can manually verify the reported abusive contents. Twitter also requests their users not to follow people if they found the content of that user as offensive. However, none of the above sites providessecurity mechanisms on server side to restrict the negative comments and the consequent damage it can bring in [4]. Unfortunately, they may be creating serious problems like online harassment, impart the sadness and loneliness, weariness, irregularities in school and decrease the performance.Different types of effort are being endorsed to reduce toxic comments given on interactive media. The researchers from the Information technology domain have been discovering the best method to determine the toxic comments automatically on interactive media. However, it is important that how it can be cease to happen. Many of the researcher have faced some common issues such as hurtful and bullied contents, a combination of offensive and poor content quality.

Since it might be solved to build application, using machine learning techniques that detect whether or not given any comment is insulting. In addition, providing full access to admin to deal it with vital factor such as he/she can hide or block these insulting comments, and also mark or flag them so that in future, site moderator can easily found it.

This paper is organized as follow: Section 2 gives the Introduction of the study. Section 3 describes the Survey of Machine learning techniques applied to the toxic comment detection. Section 4 and Section 5 gives the Toxic identification tool from two giant case study and conclusion respectively.

II. SURVEY OF MACHINE LEARNING TECHNIQUES APPLIED TO THE TOXIC COMMENT DETECTION

Machine learning is the subfield of computer science. The author Arthur Samuel defined machine learning as "computers, the ability to learn without being explicitly programmed" in 1959 [5].

Majority of Machine learning adopted in computing task where designing and programing with explicit algorithm is difficult with great accuracy; such as email filtering, detection of network intruder or malicious insider working towards a data breach [6], optical character recognition [7] and computer vision [8].

There are plenty applications of Machine Learning. The various work had been done in this direction. The major

© 2015-19, IJARCS All Rights Reserved

domains are Web Search, Computational, Finance, E-Commerce, Space exploration, Robotics, Information extraction, Social Networks, healthcare, education, agriculture, internet of things and Debugging. In the above mentioned domain, mostly the machine learning technique was applied for the prediction of the risk of coronary artery atherosclerosis [9], digital disease detection [10], dropout Prediction of students [11], ground based image analysis [12], iot [13], agriculture [14], education [15].

Here, we have given summery of the detection techniques, which are based on two machine learning algorithm supervised and unsupervised learning. However unsupervised learning algorithm was not applied fully compare to supervised learning algorithm.

Table I. **Detection of toxicity comments by machine lerning techniques**

Sr. No.	*Year*	*Authors*	*Detection of toxicity comments*	*Limitation*
1.	2009	Dawei Yin et. al.	The supervised learning was used for detecting harassment. This technique employs content features, sentiment features, and contextual features of documents with significant improvements over several baselines, including Term Frequency Inverse Document Frequency (TFIDF) approaches [16].	The experiments were done using supervised methods. The temporal or user information was not fully utilized.
2.	2012	Chen, Y. et.al.	A Lexical Syntactic Feature (LSF) architecture is used to detect offensive content and also identify the potential offensive users in interactive media. As a result, LSF framework performed significantly better than existing methods in offensive content detection as achieved precision of 98.24% and recall of 94.34% in sentence offensive detection, as well as precision of 77.9% and recall of 77.8% in user offensive detection [17].	The labeled datasets are not used in a specific domain. Able to find only 78.86% toxicity.
3.	2012	Ravi	The machine learning algorithms implemented to detect comments that may be offensive or insulting on a social networking platform. WEKA machine learning toolkit was applied and got an accuracy of 82% on the dataset [18].	The only focused was on the classifier with the highest accuracy rather than toxicity of comments.
4.	2012	Warner & Hirschberg.	In this work, the authors show a way to perform sentiment analysis in blog data by using the method of structural correspondence learning. This method accommodates the various issues with blog data such as spelling variations, script difference, pattern switching. By comparing with English and Urdu languages [19].	As a result, some constraint in mixing two languages like "bookon" in Urdu seems in English as "books" their tagger ignore such kind of offensive word.
5.	2012	Xiang el al.	The semi-supervised approach was applied for detecting profanity related offensive content on twitter using machine learning (ML) algorithms. In the experiment, the true positive rate was 75.1% over 4029 testing tweets using Logistic Regression, significantly outperforming the popular keyword matching baseline, which has a TP of 69.7%, while keeping the false positive rate (FP) at the same level as the baseline at about 3.77% [20].	The focused was on word level distribution and 860,071 Tweets. Not able to cope up with the complex feature, complex weighting mechanism and with more data.
6.	2012	Shukla et al.	A flame detector model which retrieve the written notes of the users on social networking sites and detect the flaming words and calculate the intensity level of those words [21].	The intensity of the flame is identified but not removed.
7.	2013	Dadvar et al.	An improved cyberbullying system which classifies the users' comments on YouTube using content-based, cyberbullying-specific and user-based features by applying support vector machine.[22]	Need to improve the detection accuracy for the offensive comments.
8.	2015	Razavi et al.	An automatic flame detection method, which extracts features at different conceptual levels and applies multi-level classification for flame detection [23].	The more semantic information is not extracted by preprocessing the terms and the context to which each preliminary flames were detected.

© 2015-19, IJARCS All Rights Reserved

9.	2015	Kansara &Shekokar	A framework detects only abusive text messages or images from the social network sites by applying SVM and Naïve Bayes classifiers [24].	Not able to detect audio and video which are offensive.
10.	2015	Maw Maw&Vimala A/P Balakrishnan	A system, which work through Soft Text Classifier approach using various machine learning algorithms. It is type of a screening mechanism, which alerts the users about the presence of profanity and insults. The messages are also labeled according to the subject matter [25].	While working with large volume data it may be chance to miss detection of offensive word and it affects accuracy of system.
11.	2015	Djuric et al.	A two-step method to detect hate speech using Continuous Bag of Word (CBOW) neural language model[26].	Area under the Curve very effective hate speech detectors compare to Bag of Word model in factor like time and memory.

In above table, machine learning supervised approach includes different type of decision tree algorithm, Naïve bays algorithm, Regular pattern matching algorithm, K-nearest Neighbor algorithm, novel technique and most popular and used algorithm is support vector algorithm. Most author used SVM (support vector algorithm) for classification purpose.

III. CHALLENGES

When the situation is like to handle the huge amount of data, there may be a chance to encounter the missing data points which are missed out to observe and due to which the system efficiency is affected shown by Singh, 2015 in [27]. However, Feature selection techniques can be used to resolve this problem. Taunt statement consider as no offensive words but intentionally it is an offensive. In addition, this statement or words are ignored by word-based detection system as shown in (Chen et al., 2012) [17].

Another situation like statements might have more than one meaning and different user understand different meaning of it. Since, original meaning of statement might be changed. Moreover, it can be defined as ambiguity problem [17]. Lexical resource limitation compares to English with another language also consider as Challenges.

IV. TOXIC IDENTIFICATION TOOL FROM TWO GIANT CASE STUDY

A. Google's Perspective API Built for identifying Toxic Comments.

Jigsaw and Google Counter Abuse Technology team developed one Perspective API in Conversion-AI. Machine Learning Tool used in Conversion-AI as collaborative research effort, which makes better discussions online [1].

Using Machine learning models The API create score for toxicity of an input text. Their research aims to help increase participation, quality and empathy in online conversation scale.so they consider three primary areassuch as how might machine learning methods help online conversations? What aspects of conversations can machine learning understand? Last, what are risks and challenges of using machine learning to assist online conversation? [28].

The Real Time testing done by New York Times, an early partner of Google. A devoted Staff of human's monitor over 11,000 comments per day, and only approx. 10% of articles posted have comments. Google worked with the staff and permit them to ordering thousands of comments within fraction of time, and plans to work for allowing the human moderators to police an increasing number of daily comments.

Limitation of Perspective API

Their first piece of software named as perspective is about 'toxicity which is available for English Language Only. This identifies abuse comment based on predefined set of data which is likely set by people commented in past. This model is still not perfect, as if new comments are written down which are not matched with the stored dataset then it is hard to claim as a toxic comment. Due to such limitations, more research work is required.

B. Yahoo's anti abuse AI can hunt out even the most devious online trolls:

The authors Yadav and Manwatkar had detected the offensive word using Aho-Corasick string pattern matching algorithm. The accuracy of correctly detection of offensive word is 90%and major challenge satisfied by automatically restriction of offensive word while being published in their social network prototype. For that dictionary table used to match the keyword from input text **[4]**.

In this paper, we have proposed our text filtration approach using Aho-Corasick string pattern matching algorithm for offensive word detection.

By applying the preventive measures, the restriction of offensive words is possible at real time by using the dictionary table, which is used to match the keywords from the input comments. As slang language is used for communication, the semantic relations between words are ignored.

Limitation of Yahoo

One potential way to mitigate this problem is to build a system that can detect whether or not any given comment is insulting. With such a system, website owners would have a lot of flexibility in dealing with this problem. For instance, the owner could choose to automatically block or hide these insulting comments, or flag them so that site moderators can more easily find them. The objective of this work is to do build a machine learning system that can accurately classify online posts and comments as insulting.

V. CONCLUSION

In this paper, we explored the diverse techniques and methods applied to detect the toxic comments used on social networking sites. Besides, we also outline the methods and

© 2015-19, IJARCS All Rights Reserved

limitations of very popular various machine learning techniques employed by authors to detect the toxic comments. We found that machine learning techniques are very useful and adopted by various authors due to its better performance. We identified the challenges like such as from audio/video data, the toxic word detection cannot have identified. Furthermore, more semantic word is not extracted, accuracy is not up to the mark, not able to remove the flame, language is the bar, on which future work is possible

REFERENCES

[1] Hosseini, H., Kannan, S., Zhang, B., &Poovendran, R. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv preprint arXiv:1702.08138.

[2] Duggan, M. (2014). Online Harassment. Pew Research Center.

[3] Boyd, D. (2007). Why youth (heart) social network sites: The role of networked publics in teenage social life. MacArthur foundation series on digital learning–Youth, identity, and digital media volume, 119-142.

[4] Yadav, S. H., &Manwatkar, P. M. (2015, March). An Approach for offensive text detection and prevention in social networks. In Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on (pp. 1-4). IEEE.

[5] Munoz, A. (2014). Machine Learning and Optimization. URL: https://www. cims. nyu. edu/~ munoz/files/ml_optimization. pdf [accessed 2016-03-02][Web Cite Cache ID 6fiLfZvnG].

[6] Zander, S., Nguyen, T., & Armitage, G. (2005, November). Automated traffic classification and application identification using machine learning. In Local Computer Networks, 2005. 30th Anniversary. The IEEE Conference on (pp. 250-257). IEEE.

[7] Wernick, M. N., Yang, Y., Brankov, J. G., Yourganov, G., &Strother, S. C. (2010). Machine learning in medical imaging. IEEE signal processing magazine, 27(4), 25-38.

[8] Esposito, F., &Malerba, D. (2001). Machine learning in computer vision. Applied Artificial Intelligence, 15(8), 693-705.

[9] Nikan, S., Gwadry-Sridhar, F., & Bauer, M. (2016, December). Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis. In Computational Science and Computational Intelligence (CSCI), 2016 International Conference on (pp. 34-39). IEEE.

[10] Boonchieng, E., &Duangchaemkarn, K. (2016, July). Digital disease detection: Application of machine learning in community health informatics. In Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on (pp. 1-5). IEEE.

[11] Liang, J., Li, C., & Zheng, L. (2016, August). Machine learning application in MOOCs: Dropout prediction. In Computer Science & Education (ICCSE), 2016 11th International Conference on (pp. 52-57). IEEE.

[12] Dev, S., Wen, B., Lee, Y. H., & Winkler, S. (2016). Ground-based image analysis: A tutorial on machine-learning techniques and applications. IEEE Geoscience and Remote Sensing Magazine, 4(2), 79-93.

[13] Gubbi, J., Buyya, R., Marusic, S., &Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. Future generation computer systems, 29(7), 1645-1660.

[14] Dimitriadis, S., &Goumopoulos, C. (2008, August). Applying machine learning to extract new knowledge in precision agriculture applications. In Informatics, 2008. PCI'08. Panhellenic Conference on (pp. 100-104). IEEE.

[15] Halde, R. R. (2016, September). Application of Machine Learning algorithms for betterment in education system. In Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on (pp. 1110-1114). IEEE.

[16] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB, 2, 1-7.

[17] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012, September). Detecting offensive language in interactive media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom) (pp. 71-80). IEEE.

[18] Ravi, P. (2012). Detecting Insults in Social Commentary.

[19] Mukund, S., & Srihari, R. K. (2012, June). Analyzing urdu social media for sentiments using transfer learning with controlled translations. In Proceedings of the Second Workshop on Language in Social Media (pp. 1-8). Association for Computational Linguistics.

[20] Xiang, G., Fan, B., Wang, L., Hong, J., & Rose, C. (2012, October). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 1980-1984). ACM.

[21] Shukla, S. S. P., Singh, S. P., Parande, N. S., Khare, A., & Pandey, N. K. (2012, March). Flame detector model: A prototype for detecting flames in social networking sites. In Computer Modelling and Simulation (UKSim), 2012 UKSim 14th International Conference on (pp. 553-558). IEEE.

[22] Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013, March). Improving cyberbullying detection with user context. In European Conference on Information Retrieval (pp. 693-696). Springer Berlin Heidelberg.

[23] Razavi, A. H., Inkpen, D., Uritsky, S., &Matwin, S. (2010, May). Offensive language detection using multi-level classification. In Canadian Conference on Artificial Intelligence (pp. 16-27). Springer Berlin Heidelberg.

[24] Kansara, K. B., &Shekokar, N. M. (2015). A framework for cyberbullying detection in social network. International Journal of Current Engineering and Technology, 5.

[25] Maw, M. (2016). An analysis of hateful contents detection techniques on interactive media.

[26] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., &Bhamidipati, N. (2015, May). Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web (pp. 29-30). ACM.

[27] https://conversationai.github.io/

[28] Rohan Shetty, Kalyani Nair, Shivani Singh, Shantanu Nakhare, GopalUpadhye (2015), A System To Detect Inappropriate Messages In Online Social Networks, International Journal of Advanced Computational Engineering and Networking, (40-43).

© 2015-19, IJARCS All Rights Reserved