A Survey on Toxic Comment Classification

Abhishek V. Lokhande[1] Prof. Dr. B. K. Sarkar[2]
[1,2]Department of Computer Engineering
[1,2]P.V.P.I.T Savitribai Phule Pune University, Pune, India

Abstract— The growing demand for the Internet has given rise to social issues such as abusive behaviour which comprises intolerable comments, personal attacks, online harassment, and bullying. It is indispensable to categorize the comments based on toxicity to prevent the bullying on the social network. The industry, as well as the research community including the Jigsaw team from Google, has started putting endeavours towards addressing this belligerent issue. Google is looking for proposing an efficient model for online prediction and scoring of toxic comments. In this paper, we are presenting a comprehensive survey on a classification of toxic comments and also providing a future roadmap for online toxic comments classification.

Keywords: Deep Learning, NLP, Toxic Comment Classification, Machine Learning, Text Mining

I. Introduction

The usage of social media websites is increasing at a rapid rate. A few examples of online media are online news, e-commerce websites and gaming sites. The rise in usage of online social platforms increases communication among people worldwide. The communication takes place in the form of comments, review, and feedback which may be positive or negative. The positive discussion is always healthy for any community but the negative and toxic comments can leave a remarkable bad impact on the society. The toxic comments are commonly defined as any communication that mocks a person or a group on the basis of some characteristic such as colour, ethnicity, race, gender, sexual orientation, religion, nationality or any other characteristic [6]. The survey report by McAfee [7] states that almost half of the total Indian Youth have experienced online harassment trolls and hatred comments on social media. In [7], the author has revealed that the common reason behind the number of suicides committed by teens is toxicity in the comments. The impact of toxic comments has given rise to frustration, mental stress and also leads to trauma in certain cases. In a recent survey, 39% of experts in this field of online discourse have claimed that the future of online communication would be full of harassment and trolls [9].

Some of the social platforms already use some mechanism that enables the users to report the contents as toxic and can be removed from social media platforms [8]. The social media organizations such as Twitter, Facebook, Instagram has a number of employees working on monitoring the social media content, which is being uploaded daily in form of reviews, comments and opinion. If the comments are found as inappropriate or abusive then respective action is taken. As a consequence, the comment could be deleted or the user could be blocked from the platform.

Over the last decade, the interest in toxic comment detection and particularly automation of the process has been increased. Many researchers have used NLP (natural language processing) and other machine learning tools for the identification of the toxic comments. The researchers and application developers from all over the world are working on to reduce the negative impact of toxic comments on society. Although there are many challenges to create a permanent and effective solution to the problem but still many organizations such as Jigsaw from Google are putting efforts to provide the automatic solution to this problem.

The problem could be solved by developing an application using machine learning techniques that can identify the toxicity level of the comment and also classify the comment in subcategories. In addition to this, the user should be able to hide or block these toxic comments, and also able to mark them so that the website administrator can easily detect it for future reference.

In this paragraph, we are elaborating the organization of this paper. The first section discusses about background details, the second section of the paper provides information on the existing and related works on toxic comments detection and classification. The third section concludes the paper and also provides directions for future work.

II. Related Work

The classification of the toxic comments has been intensively researched in past few years, mostly in the social media context where researchers have applied different machine learning algorithms to classify the online comments into various toxic classes. We are providing a survey on the existing machine learning techniques for classification of toxic comments in this section.

In [1], the authors have proposed a technique that uses a recurrent neural network (RNN) for classifying toxic comments into multiple classes i.e. toxic, severe toxic, obscene, threat, insult or hate. They have conducted the experiment using both non-neural based model which uses simple TFIDF(term frequency–inverse document frequency) sentence vector and neural network based model which uses LSTM(Long Short Term Memory) RNN with pre-trained word embedding. Both the models performed very well on the provided dataset but considering the complexity of the provided dataset where comments contain highly non-standard vocabulary and the number of positive examples was low, neural network performed slightly better than non-neural based model. Future scope for the improvement, in this case, would be using word embedding with more attention on a particular word in a particular comment. It can enhance the accuracy in classification of the toxic comments.

In paper [2], authors have used Convolutional Neural Networks (CNN) over the traditional Bag of words (BoW) text classification technique. For this experiment, they have used Wikipedia dataset provided by Kaggle competition. The dataset contains comments from Wikipedia’s talk page edits which have been labeled by human raters for toxic behavior. And the classification is done in six type of toxic classes i.e. toxic, severe toxic, obscene, threat, insult, identity hate. For the given dataset, BoW model is used with different machine learning classification methods namely Support Vector Machines (SVM), Naive Bayes (NB), k-Nearest Neighbor (kNN) and Linear Discriminate Analysis (LDA). After applying both the techniques, F1 score is calculated and the result shows that the use of CNN with word embedding has outperformed the traditional BoW text classification model which uses SVM, kNN, NB and LDA methods. It has achieved accuracy over 90% while another traditional algorithms have accuracy between 65 to 85 %.

In paper [3] authors have used various deep learning approaches for classifying online comments into different toxic classes. Authors have considered both type of classification, binary as well as the multi-label classification. In binary classification, they have only one class where the given comment is classified as toxic or non-toxic whereas in multi-class classification, the comments are classifies into several toxic classes. In this research work, the authors have studied and compared the effect of three different kinds of a neural network, i.e. Multilayer perceptron (MLP), Long-short-term memory (LSTM) and Convolutional Neural Networks (CNN) with two level of granularity (word level and character level). It is shown that LSTM shows more accuracy for word-level binary and MLP for multi-label classification and for character level binary classification, CNN has the highest level of accuracy.

In paper [4], authors have used two approaches to achieve the accuracy in classification. The neural network based approach and the non-neural network-based approach for classification. The Wikipedia’s talk page edits dataset is used in this experimental study. For non-neural based approach, authors have used Naïve Bayes algorithm combined with logistic regression. In terms of accuracy, the result of the classification for non-neural based approach is very good but the F1 score for this technique is very low. On the other hand, neural network based model(RNN stacked and bidirectional) performed better than non-neural based model in terms of accuracy as well as in F1 score.

In paper [5], author has used deep learning method i.e. Convolutional Neural Networks (CNN) for text classification. The dataset used in this experimental study is Twitter dataset which contains thousands of tweets from different users and it has been labeled with three classes, i.e. hate, offensive language, and neither. After applying CNN on the twitter dataset, it classifies these tweets into the three classes as hate, offensive language, and neither. It is found that after applying simple CNN algorithm, the accuracy is 91% in terms of classification but this model has misclassified some of the non-toxic tweets as toxic tweets. In the future scope of this paper, the accuracy can be increased by doing error analysis, collecting more data so that the model will get trained with different types of toxic classes.

In article [11], author has used Keras library to classify the text into multiple classes. Keras is an open source neural network library written in Python. In order to classify the sentence into multiple classes, the author has first represented the sentence into the bag of words model and then used a multi-layer neural network to train the model. In this work, the author has got the results which are satisfactory. The main focus of the research was to create a strong baseline for further research.

In [12], the author has applied machine learning techniques to perform automated offensive language detection. The study mainly focuses on two categories 'sexual' and 'racist'. The data that used for detection and classification is originating from the Dutch distribution of the social network Netlog, which contains over seven million blog messages. The author has implemented two supervised learning methods Naive Bayes and Support Vector Machine which is used to train the model for classification. In order to build such training set offensive messages should be efficiently extracted out of the corpus and to achieve this, an information retrieval system, expanded with a Rocchio query expansion technique, is applied. The model is first trained on the labeled set and afterwards the performance is tested on the validation set. The Naive Bayes classifier does not perform well on the validation set and is therefore disregarded in the further analysis. Support Vector Machine implementation achieves results of approximately 69% precision and 62% recall. However, these results are obtained by ignoring very small messages, since SVM has difficulties classifying messages that do not contain much information. A more reliable but less dynamic method is designed to tackle such issues based on word lists. This method named as a semantic classifier obtains reasonable results on the validation set with a recall of 93% but more importantly is highly complementary with the SVM classifier. This method outperforms all others and achieves a precision of 100% combined with a 79% recall.

In paper [15], the authors have examined different methods to detect hate speech in social media while distinguishing this from general profanity. Their aim is to establish lexical baselines for this task by applying supervised classification methods. The dataset used to conduct this research contains English tweets annotated with three labels: hate speech (HATE), offensive language but no hate speech (OFFENSIVE) and no offensive content (OK). In this paper, authors have applied a linear Support Vector Machine (SVM) classifier and used three groups of features extracted for these experiments: surface n-grams, word skip-grams, and Brown clusters to distinguish between hate speech, profanity, and other texts. The best result was obtained by a character 4-gram model achieving 78% accuracy. The results presented in this paper showed that distinguishing profanity from hate speech is a very challenging task.

In paper [17] authors have given more focus on the issue that the hate speech data is unbalanced in nature and it lacks unique, discriminative features, therefore is found in the ‘long tail’ in a dataset that is difficult to discover. The research model is evaluated on the largest collection of hate speech datasets based on Twitter. Authors have proposed Deep Neural Network (DNN) structures that are empirically shown to be a very effective feature extractor for identifying specific types of hate speech. These include two DNN models inspired and adapted from the literature of other Machine Learning tasks: one that simulates skip-gram like feature extraction based on modified Convolutional Neural Networks (CNN), and another that extracts orderly information between features using Gated Recurrent Unit networks (GRU). By analyzing the result we can conclude that authors are able to outperform state of the art by up to 6% points in macroaverage F1, or 9 % points in the more challenging case of identifying hateful content.

As of now, research studies have been only focusing on individual comments to detect offensive text, but in paper [22], authors have taken user's context into accounts such as users' characteristics and profile information to improve the cyberbullying detection accuracy. The dataset used for this study is collected from YouTube movies comments. For each comment, the user id, its date and time have also been stored. Only the users with public profiles (78%) are kept. The final dataset consists of 4626 comments from 3858 distinct users. The comments are manually labeled as bullying (9.7%) and non-bullying based on the definition of cyberbullying. For classification purpose 3 features are extracted from the comments, 1)Content-based features- These features are based on the contents of the comments itself and are frequently used for sentiment analysis, 2)Cyberbullying features - The second set of features aims at identifying frequent bullying topics such as minority races, religions and physical characteristics and 3) User-based features - To be able to exploit information about the background of the users in the detection process like age of the user. Authors have used these incremental features to train the Support Vector Machine to classify comments as bullying or non-bullying. The results showed that incorporation of context in the form of users' activity histories improves cyberbullying detection accuracy. In the future scope of this work, one can develop a model that detects expressions involving sarcasm or implicit harassment.

Sr. No.	Title of Research Paper	Authors	Technique Used	Findings
1	Mining Offensive Language on Social Media[24]	Alessandro Maisto, Serena Pelosi, Simonetta Vietri, Pierluigi Vitale	Used lexicon-based method for automatic identification and classification of taboo expressions.	Using lexicon-based method for automatic identification helps in identifying the toxic words in local languages. All the comments then classified into 4 toxic classes.
2	Abusive Language Detection in Online User Content [10]	Chikashi Nobata, Joel Tetreault, Achint Thomas, Yi Chang	Used supervised classification method with NLP features for detection and classification purpose	Results show that using character n-grams alone worked very well in noisy data sets and also outperforms a deep learning based model.
3	Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter[19]	Z. Waseem and D. Hovy.	Character n-gram based approach is used	Through this research authors have provided a list of criteria based in critical race theory to identify racist and sexist slurs, which will help other researches in this field.
4	Ex Machina: Personal Attacks Seen at Scale[16]	L. D. Ellery Wulczyn, Nithum Thain	Used crowdsourcing and machine learning to analyze personal attacks	Results shows that the classification done by the automated system is as good as work done by manual user using crowdsourcing.

Table 1: Survey Table

III. CONCLUSION AND FUTURE WORK

In this paper, we explored different machine learning techniques used to detect and classify toxic comments on social media platforms. We also outlined the limitations of these algorithms. We have found that machine learning is the key to the classification of text and applying predictions to detect the toxicity level of the text. There are many areas where we can improve in terms of accuracy in finding the toxic comment and its toxicity level.

In future work, we can focus on performance and error analysis of the model as lots of comments are misclassified into the hate category. Previous work has achieved success using various algorithms on data in English language but in future, we can consider having data in regional languages. We can also work on after work of the detection of the toxic comments like automatic blocking of the user, auto-deletion of harmful comments on social media platforms.

REFERENCES

[1] Manav Kholi, Emily Kuehler and John Palowitch "Paying attention to toxic comments online", Stanford University journal Year 2017

[2] Spiros V. Georgakopoulos et al. "Convolutional Neural Networks for Toxic Comment Classification", Cornell University arXiv:1802.09957 Year 2018

[3] Kevin Khieu and Neha Narwal, "Detecting and classifying toxic comments", Stanford University journal CS224N [2017].

[4] Maxime Rivet and Mael Tran, "Toxic comments classification", Stanford University journal Year [2016].

[5] Shanita Biere "Hate Speech Detection using Natural language processing techniques", VU-Vrije University Amsterdam journal [2018]

[6] John T. Nockleby. "Hate Speech. In Leonard W. Levy, Kenneth L. Karst, and Dennis J. Mahoney, editors, Encyclopedia of the American Constitution", pp. 1277–1279. Macmillan, 2nd edition, Year 2000.

[7] Duggan, M. "Online Harassment". Pew Research Center Internet & Technology, [2017]

http://www.pewinternet.org/2017/07/11/online-harassment-2017

[8] Yadav, S. H., &Manwatkar, P. M. “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, [2015]

[9] L.Rainee, J.Anderson, and J.Albright, “The future of free speech, trolls, anonymity and fake news online”, Pew Research Center Internet & Technology [2017] http://www.pewinternet.org/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online

[10] Yi Chang & al. “Abusive language detection in online user content” www`16 Proceeding of the 25th International Conference on World Wide Web Page 145-153. [2016]

[11] T. Sterbak, “A strong baseline to classify toxic comments on Wikipedia with fasttext in keras” [2018] https://www.depends-on-the-definition.com/classify-toxic-comments-on-wikipedia

[12] Baptist Vandersmissen, "Automated detection of offensive language behavior on social networking sites". [2012] https://lib.ugent.be/catalog/rug01:001887239

[13] Spiros V. Georgakopoulos et al., “Convolutional neural networks for toxic Comment classification” [2018] arXiv:1802.09957

[14] S. Jogeklar, “First time with kaggle: A convnet to classify toxic comments with keras” [2018] https://medium.com/@srjoglekar246/first-time-with-kaggle-a-convnet-to-classify-toxic-comments-with-keras-ef84b6d18328

[15] Shervin Malmasi, Marcos Zampieri, "Detecting Hate Speech in Social Media". [2017] arXiv:1712.06427

[16] L. D. Ellery Wulczyn, Nithum Thain, “Ex-machina: Personal attacks seen at scale” arXiv:1610.0891 [2016]

[17] Ziqi Zhang. “Hate speech detection: A solved problem? the challenging case of long tail on twitter” arXiv:1803.03662 [2018]

[18] H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths. A web of hate: Tackling hateful speech in online social spaces. arXiv:1709.10159 In TA-COS [2016]

[19] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of NAACL-HLT, pages 88–93, [2016]

[20] M Ramakrishna Murty, JVR Murthy, and Prasad Reddy “Text Document Classification based-on Least Square Support Vector Machines with Singular Value Decomposition.” International Journal of Computer Applications 27, 7 [2011]

[21] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. “Detecting offensive language in interactive media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT)”, International Conference on and 2012 International Conference on Social Computing.[2012]

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

[23] E. Wulczyn, N. Thain, and L. Dixon. Wikipedia talk labels: Personal attacks. 2 [2017].

[24] Alessandro Maisto, Serena Pelosi, Simonetta Vietri, Pierluigi Vitale “Mining Offensive Language on Social Media”[2017]