The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 9 Issue 1, 2021
pp: 44 - 50

Toxic Comments Classification in Social Networking

[1]C. Lakshmi, [2]D. Ananya, [3]M. Sreevani, [4]M. Sreedevi
[1,2,3,4] UG Student, Department of CSE,
Dr K V Subba Reddy College Of Engineering For Women, Kurnool, Andhra Pradesh, India

Abstract

Today, users post a lot of comments on new portals, social networks, and forms. The majority of systems use some kind of automatic discovery of toxicity using machine learning models because it is impossible to manually moderate all of the comments. In this work, we performed a systematic review of the state-of-the-art in toxic comment classification using machine learning methods. We gathered information from 31 primary relevant studies. First, we looked into when and where the papers were written, as well as their maturity level. Every primary study's evaluation metric, used machine learning techniques, toxicity classes, and comment language were examined in our analysis. We conclude our work with a comprehensive list of gas currently being studied and suggestions for future research topics on the online toxic comment classification issue

1. Introduction

Over the years, social media and social networking use have been increasing exponentially due to an upsurge in the use of the internet. Flood of information arises from online conversation in a daily basis as people are able to discuss, express themselves and air their opinion via these platforms. While this situation is highly productive and could contribute significantly to the quality of human life, it could also be destructive and enormously dangerous.

While discussion or a conversation is opened, it is quite obvious that debates may arise due to differences in opinion. But often these debates take a dirty side and may result in fights over the social media during which offensive language termed as toxic comments may be used from one side. These toxic comments may be threatening, obscene, insulting or identity-based hatred. So, these clearly pose the threat of abuse and harassment online. Consequently, some people stop giving their opinions or give up seeking different opinions which result in unhealthy and unfair discussion.

As a result, different platforms and communities find it very difficult to facilitate fair conversation and are often forced to either limit user comments or get dissolved by shutting down user comments completely. This study focuses on building a multi-headed model to detect different types of toxicity like threats, obscenity, insults, and identity-based hate. Detecting and controlling verbal abuse in an automated fashion is inherently a natural language processing task. Natural Language Processing, (NLP), is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 9 Issue 1, 2021
pp: 44 - 50

the human languages in a manner that is valuable.

Most NLP techniques rely on machine learning to derive meaning from human languages. Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model for example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions. Toxic comment classification on online channels is conventionally carried out either by moderators or with the help of text classification tools. With recent advances in Deep Learning (DL) techniques, researchers are exploring if DL can be used for comment classification task.

Text classification is a classic topic for natural language processing and an essential component in many applications, such as web searching, information filtering, topic categorization and sentiment analysis. Text transformation is the very first step in any form of text classification. The online comments are generally in non-standard English and contain lots of spelling mistakes partly because of typos (resulting from small screens of the mobile devices) but more importantly because of the deliberate attempt to write the abusive comments in creative ways to dodge the automatic filters.

Nowadays, the flow of data over the internet has grown dramatically, especially with the appearance of social networking sites. Social networks sometimes become a place for threats, insults, and other components of cyberbullying. A huge number of people are involved in online social networks.

Toxic comments are textual comments with threats, insults, obscene, racism, etc. In recent years there have been many cases in which authorities have arrested some users of social sites because of the negative (abusive) content of their personal pages. Hence, the protection of network users from anti-social behavior is an important activity. One of the major tasks of such activity is automated detecting the toxic comments.

Architecture:

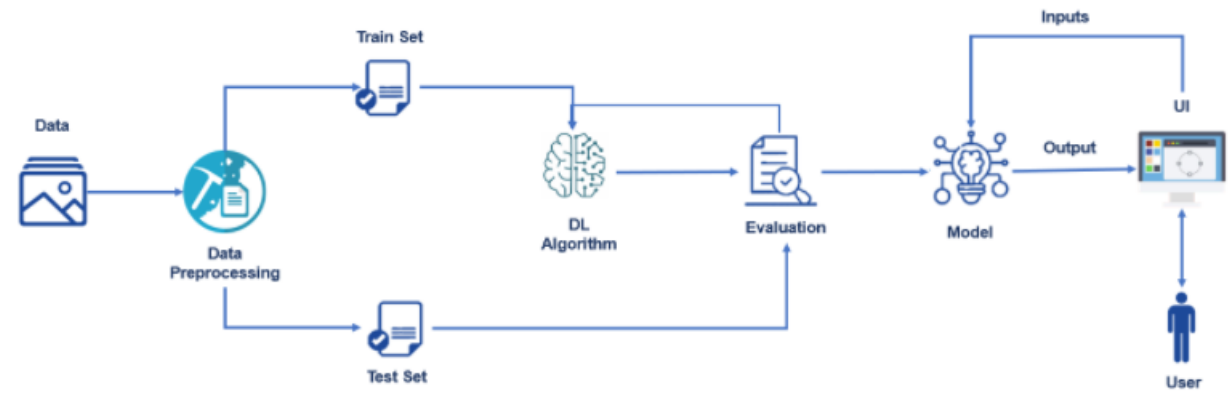

Fig.1 Architecture

2. Literature Review

Bag of words statistics and bag of symbols statistics are the typical source information

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 9 Issue 1, 2021
pp: 44 - 50

for the toxic comments detection. Usually, the following statistics-based features are used: length of the comment, number of capital letters, number of exclamation marks, number of question marks, number of spelling errors, number of tokens with non-alphabet symbols, number of abusive, aggressive, and threatening words in the comment, etc. A neural network model is used to classify the comments.

Nguyen and Nguyen[22] proposed a model for sentiment label distribution that involved a combination of using a deep CNN for character -level embedding(in order to increase information for word-level embedding's) with a bidirectional LSTM which produced sentence-wide feature representation from word-level embeddings.This model attained a best prediction accuracy of 86.63%on the Stanford twitter sentiment corpus. They are findings indicate the prospective advantages of utilizing LSTM and deep CNN models on the task of toxicity classification.

Yu and Wang [26] proposed a word vector refinement model that could be applied to pertained word vectors (example Word2vec and Glove) to enhance sentiment information capture. The refinement model was based on adjusting vector representations of words so that they could be closer to both semantically and sentimentally similar words and farther away from sentimentally dissimilar words. The word embedding from the refined model (Re), Re (Word2vec) and Re(GloVe) respectively, improved Word2Vec and GloVe by 1.7% and 1.5% averaged overall classifiers for binary classification, and both improved by 1.6% for fine-grain classification .These result suggest potential further improvement of our model may result from addition enhancement of word vectors(from using this refinement model) that capture more semantic information.

Chu and Jue [30] compared the performance of various deep learning approaches to this problem, specifically using both word and character embeddings. They assessed the performance of recurrent neural networks with LSTM and word embeddings, a CNN with word embeddings, and a CNN with character embeddings. The best performance they achieved was a 93% accuracy using the character level CNN model.

We further extend the previous analysis to assess performance of deep learn approaches :Forward LSTM , bi-directional LSTM, CNN , Multilayer perceptron networks, and support vector machines , both at the word and character level and for binary and multi-label classification.

This study has been carried out using the systematic literature review (SLR) methodology described in [2]. First, we have denied the SLR protocol. Then, we performed the study selection and the data extraction process whose outcome is the final list of papers. The main steps of the SLR protocol are listed and elaborated in the next subsections.

The characteristics of the data collected for this study is analyzed in this section. This consists of data collected using Jigsaw. A dataset of comments from Wikipedia's talk page edits is also used. Jigsaw analyses Wikipedia comments (i.e. either toxic or non-toxic), and make the dataset available for those who want to further work on the research. The contribution of Jigsaw is to develop and illustrate a method that combines crowd sourcing and machine learning to analyze personal attacks. This section also discuss text mining and, also the

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 9 Issue 1, 2021
pp: 44 - 50

processing of text carried out using the term frequency-inverse document frequency (TF-IDF) technique

```
#Define the parameters /arguments for ImageDataGenerator class
train_datagen=ImageDataGenerator(rescale=1./255,shear_range=0.2,
        rotation_range=180,zoom_range=0.2,horizontal_flip=True)

test_datagen=ImageDataGenerator(rescale=1./255)
```

Fig.2 Parameter

3. Proposed System

The data we used came from Jigsaw/Google's Kaggle challenge to classify toxic comments. There are 159,571 examples of Wikipedia comments that have been labeled as toxic behavior by human raters in the dataset. The labels for toxic, severe-Toxic, obscene, threat, insult, and identity-Hate are all Boolean labels, and the data comes in the form of "id, comment Text, toxic, severe-Toxic, obscene, threat, insult, identity-Hate>." For supervised machine learning, the majority of primary studies have used one or two data sets with toxic comments. Jigsaw's data set for the Toxic Comment Classification Challenge competition, which is hosted on Kaggle, is the most frequently used [1]. Twenty-two selected primary studies make use of it. Jigsaw Unintended Bias in Toxicity Classification Kaggle's competition later updated this data set. Numerous Wikipedia comments from the aforementioned data set have been categorized as toxic by human raters. Toxicity can be classified as: obscene, extremely obscene, toxic, threatening, insulting, and identity hatred. Other data sets that are used are specific to a particular primary study. Third parties have created some of these additional data sets: Davidson's Twitter dataset, Hosseinmardi et al.'s Instagram dataset, Task 1 of Semeval2018 included nearly 7000 tweets, the Twitter Hate Speech dataset, a dataset of tweets sent by members of the United States House of Representatives, the Wikipedia Detox corpus, and live video game chat conversations. The authors of these studies created the remaining datasets used in primary studies: reviews from Udemy, synthetic training data from Facebook comments posted in response to popular news articles, a large collection of over 104 million Reddit comments, a dataset from Facebook page posts, 9.4 thousand manually labeled entertainment news comments for identifying Korean toxic speech, comments in Hindi and English Machine Learning Methods for Toxic Comment Classification 211 both scraped from Facebook and Twitter, and custom-prepared data from Facebook, Twitter, Instagram, and Whatsapp in Arabic

For our binary classification task, we used the given toxic label as our binary indicator for whether or not a given comment was toxic. Data preprocessing was done with R and Python. We were able to obtain a total of 30,590 examples for use in training, validation, and testing after balancing the dataset through random subsampling. We divided our data 60-20-20 to separate them. We used the same balanced dataset as for the binary

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 9 Issue 1, 2021
pp: 44 - 50

classification task for our multi-label classification task, but we also included labels for each specific type of toxicity (severe toxic, obscene, threat, insult, and identity-Hate). In order to reduce the number of 0-labels in this task, we also experimented with a version of our dataset that had 0 non-toxic sentences (15,295 examples) to try to further balance our labels.

Our word count; analysis represented each word in our dataset as a vector using GloVe word vectors that had been trained from Wikipedia data. Using Keras' "Embedding Layer package," we trained our own character-level embeddings for our character-level analysis. Comment padding and clipping were also used in our LSTM and convolutional neural network models to guarantee that each example in our dataset had the same number of words and characters. By treating each input comment equally, we were able to batch process the data effectively lengths

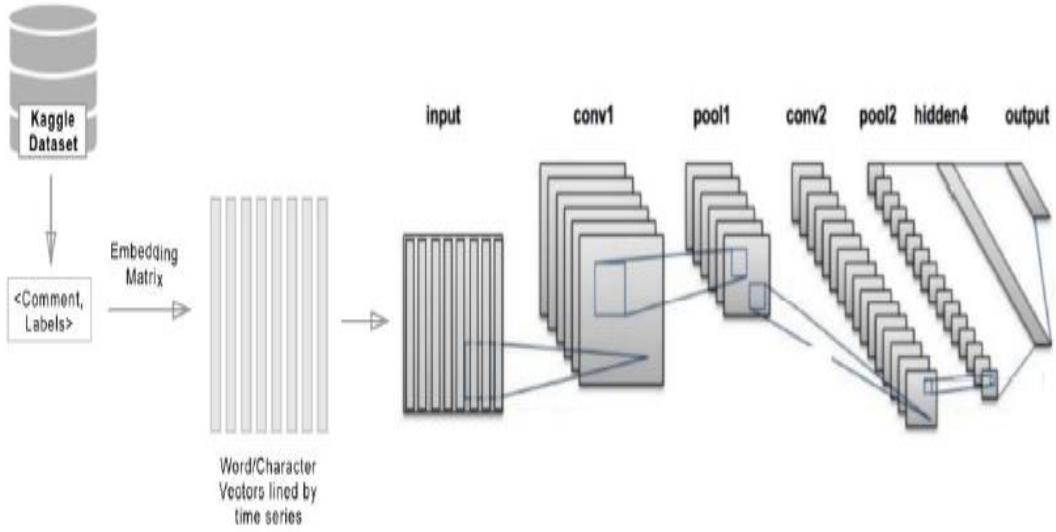

Fig.3 Proposed Method

4. Conclusion

Communication is one of the basic necessities of everyone's life. People need to talk and interact with one another to express what they think. Over the years, social media and social networking have been increasing exponentially due to an upsurge (rise) in the use of the internet. Flood of information arises from online conversation on a daily basis, as people are able to discuss, express themselves and express their opinions through these platforms. While this situation is highly productive and could contribute significantly to the quality of human life, it could also be destructive and enormously dangerous. The responsibility lies on the social media administration or the host organization to control and monitor these comments. This research work focuses on developing a model that would automatically classify a comment as either toxic or non-toxic using logistic regression. Therefore this study aim to develop a multi-headed model to detect different types of toxicity like threats, obscenity, insults, and identity based-hate. By collecting and preprocessing toxicity classified comments for training and testing using term frequency-inverse frequency document (TF_IDF)

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 9 Issue 1, 2021
pp: 44 - 50

algorithm, developing a multi-headed model will detect different types of toxicity using logistic regression to train the dataset and to evaluate model using confusion metrics. Our best models with regards to word-level binary and multi-label classification were our LSTM and CNN Models. For character-level binary classification, our CNN models yielded the best performance, although overall our word-level models out performed our character-level models. Additionally, we did try layering a character-level CNN with a bidirectional LSTM as suggested by NGUYEN and NGUYEN, but we are unable to recreate the high accuracy metrics they attained in their work.

References

1. A. Vaidya, F. Mai, Y. Ning, Empirical Analysis of Multi-Task Learning for Reducing Model Bias in Toxic Comment Detection, ArXiv190909758 Cs, Mar. 2020, Accessed: Jul. 03, 2020. http://arxiv.org/abs/1909.09758.
2. A. Bleiweiss, LSTM neural networks for transfer learning in online moderation of abuse context, ICAART 2019 - Proceedings of the 11th International Conference on Agents and Artificial Intelligence, Prague, Czech Republic,2019.
3. A. P. Patil, A. Mohammed, G. Elachitaya, M. Tiwary, Practical Significance of GA PartCC in Multi-Label Classification, Proceedings of the 2019 Ieee Region 10 Conference (tencon 2019): Technology, Knowledge, and Society, Kerala, India,2019.
4. A. Elnaggar, B. Waltl, I. Glaser, J. Landthaler, E. Scepankova, F. Matthes, Stop Illegal Comments: A Multi-Task Deep Learning Approach, ACM International Conference Proceeding Series, 2018.
5. A. G. D'Sa, I. Illina, D. Fohr, Towards non-toxic landscapes: Automatic toxic comment detection using DNN, ArXiv191108395 Cs Stat, Nov. 2019, Accessed: Jul. 03, 2020. http://arxiv.org/abs/1911.08395.
6. B. Kitchenham, S. Charters, Guidelines for performing Systematic Literatur Reviews in Software Engineering(2007).)
7. B. van Aken, J. Risch, R. Krestel, A. L• oser, Challenges for Toxic Comment Classification: An In-Depth Error Analysis, ArXiv180907572 Cs, Sep. 2018, Accessed:Jul. 03, 2020. http://arxiv.org/abs/1809.07572.
8. Chu,T&Jue,K.Comment Abuse Classification with Deep Learning
9. https://smartinternz.com/guided-project/intelligent-alert-system-for-forest-tribal-people-duplicate-
10. https://in.linkedin.com/in/mansi-ag-41420a1ab#:~:text=Intelligent%20Alert%20System%20For%20Forest%20Tribals&text=This%20system%20will%20monitor%20the,the%20image%20of%20the%20animal%20.
11. https://www.google.com/search?source=univ&tbm=isch&q=Intelligent+alert+system+for+forest+tribal+people&sa=X&ved=2ahUKEwif8sv7r6PxAhWw63MBHQP_A-oQjJkEegQIGxAC&biw=1517&bih=631
12. https://www.google.com/url?sa=i&url=https%3A%2F%2Fsmartinternz.com%2Fguided-project%2Fintelligent-alert-system-for-forest-tribal-people-duplicate-

&psig=AOvVaw1o_vagI79XQ9Et12d6I45k&ust=1624181247367000&source=images&cd=vfe&ved=0CAcQjRxqFwoTCLDHhZaxo_ECFQAAAAAdAAAAABAJ
13. https://news.mongabay.com/2020/09/new-artificial-intelligence-could-save-both-elephant-and-human-lives/
14. https://keras.io/api/preprocessing/image/#imagedatasetfromdirectory-function
15. https://www.youtube.com/watch?v=YNKo11c3EX0
16. https://www.google.com/search?q=intelligent+alert+system+for+forest+tribal+people+document+copy&sxsrf=ALeKk033tKgOLVzotSC3Z-ShtBEBMg-A9w:1625155201086&tbm=isch&source=iu&ictx=1&fir=MCRHJ4N_-o0P8M%252C5hToV5wCrzI8cM%252C_&vet=1&usg=AI4_-kQhzELrTepg9LorVuJ15E_bgbx60Q&sa=X&ved=2ahUKEwiRttzznsLxAhWTILcAHbI2BNMQ9QF6BAgUEAE#imgrc=MCRHJ4N_-o0P8M
17. https://www.google.com/search?q=intelligent+alert+system+for+forest+tribal+people+document+copy&sxsrf=ALeKk033tKgOLVzotSC3Z-ShtBEBMg-A9w:1625155201086&tbm=isch&source=iu&ictx=1&fir=MCRHJ4N_-o0P8M%252C5hToV5wCrzI8cM%252C_&vet=1&usg=AI4_-kQhzELrTepg9LorVuJ15E_bgbx60Q&sa=X&ved=2ahUKEwiRttzznsLxAhWTILcAHbI2BNMQ9QF6BAgUEAE#imgrc=_KNuXaDqyb76qM
18. https://www.google.com/search?q=forest+tribe+pepole+documentary&oq=forest+tribal+pepole+docum&aqs=chrome.1.69i57j33i22i29i30.12489j0j7&sourceid=chrome&ie=UTF-8