International Journal of Research in Engineering, Science and Management
Volume 5, Issue 4, April 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

A Study on Foul Language Detection

K. Nartkannai[1], P. Preethi[2], M. Rishitha[3], V. Sai Vaishnavi[4], K. Aarti Chowdary[5*]

[1]*Assistant Professor, Department of Computer Science and Engineering, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India*

[2,3,4,5]*UG Student, Department of Computer Science and Engineering, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India*

***Abstract*: There is always a substantial risk of scorn and even harassment when one engages in online activity, whether on message board discussions, comments, or social media. Words that are inappropriate are unfortunately frequent online and can have a significant impact on a community's civility or a user's experience. To fight abusive language, many websites have standards and guidelines that users must follow, as well as human editors who work in tandem with systems that utilize regular expressions and blacklists to capture foul language and so remove a post. The demand for high-quality automatic abusive language classifiers is growing as individuals speak more online. As this is the complex problem ML is being suggested as an effective tool to detect Abuse language. Here is the detailed analysis of the existing systems comparing their methodology and accuracy.**

***Keywords*: Abusive language, Machine Learning.**

1. Introduction

On the internet today, there are a variety of online discussion forums where users may express, discuss, and trade their thoughts and opinions on a variety of subjects. Users can share their opinions in the form of comments on news portals, blogs, and social media channels such as YouTube, Instagram, Twitter, Facebook etc. It has been noted that user talks in such forums frequently divert and become inappropriate, such as screaming abuses, passing nasty and discourteous comments on people or certain groups/communities.

The intent of a given textual information is defined as inappropriate if it is rude or discourteous toward certain individuals or groups of individuals, to cause or capable of causing harm (to oneself or others), related to an activity that is illegal under the laws of the country, or has extreme violence.

Similarly, it has been discovered that certain virtual agents or bots respond to users with offensive messages. As a result, unwanted messages or comments are slowly eroding the effectiveness of user experiences online. As a result, automatic detection and filtering of such offensive language has become a significant concern in terms of increasing the quality of user chats.

2. Literature Survey

Offensive Language Detection using Multi-Level Classification [1] which is proposed by Amir H. Razavi, Diana Inkpen, Sasha Uritsky and Stan Matwin processed the data in three levels. They have set up Insulting and Abusing Language Dictionary (IALD). IALD has the patterns of flames (i.e., abusive or offensive sentences which are preprocessed). In the first level of classification, they have used Complement Naïve Bayes classifier. Then in the level, proceeded with Multinomial Updatable Naïve Bayes classifier and lastly picked up a rule-based classifier named Decision Table/Naive Bayes hybrid classifier. All the three levels classify the messages/sentences as Flame/Okay. In each step, having a total of 1525 messages of which 1038 where non-abusive or Okay and 487 were abusive or Flame, 10% of the data was considered for testing. The dataset is the fusion of two data sources that are explicitly formatted.

Detecting Offensive Language in social media to Protect Adolescent Online Safety [2] has used Lexical Syntactic Feature (LSF) architecture which is discovered by Ying Chen, Sencun Zhu, Yilu Zhou, Heng Xu. Data set is sourced from 2,175,474 users that commented under top 18 videos under different sections of YouTube. Naïve Bayes and SVM are used to classify the comments and 10-fold cross validation is used to check the efficiency. The LSF approach has to tolerance to misspellings and informal sentences and improved the traditional approach by binding it with style and structure of the sentence with the lexical features.

Azalden Alakrot, Liam Murray, Nikola S. Nikolov have worked Towards Accurate Detection of Offensive Language in Online Communication in Arabic [3] has incorporated three approaches in their research. First, without data preprocessing and then routed towards second approach in two ways i.e., without stemming and with stemming with data-preprocessing where the precision is increasing at each step. Now, in the last step using SVM and also combining N-gram features with stemming which increased the recall. Thus, it is not recommended to use both of them on the same machine learning model. The extra normalization and N-gram feature has been beneficial. The dataset is built of 15,050 comments in Arab under YouTube videos.

Abusive Language Detection in Online Conversations by Combining Content- and Graph- Based Features [4] which is proposed by Noe Cecillon, Vincent Labatut, Richard, Dufour and Georges Linares. They proposed fusion methods integrating content- and graph-based features. In content- based

*Corresponding author: aarti.c.kolasani@gmail.com method raw data is fed as input to Support Vector Machine (SVM) classifier to distinguish abusive and non-abusive messages. In Graph-based method it completely ignores the content messages and mainly focuses on the dynamics of the conversation, it is a three-stepped method (1) extracting a conversational graph, (2) computing topological measures, (3) train to SVM to distinguish between abusive and non-abusive messages.

Offensive Language Detection Explained [5] by Julian Risch, Robin Ruff, Ralf Krestel has given us the analysis and comparison of 4 methods. They are Naive Bayes, local interpretable model-agnostic explanations, Layer-wise relevance propagation and long short-term memory (LSTM) neural network. The key points were computational power and unfair advantages on deletion of foul language messages, relevance and explanative power

Detecting Flames and Insults in Text [6] which is discovered by Altaf Mahmud, Kazi Zubair Ahmed, Mumit Khan. They proposed a new approach for an automated system to differentiate between information and personal attacks which may contains abusive expressions, racism, neutral in a given data. In linguistics, abusive messages are viewed as an extreme subset of the subjective language of a given sentence in order to extract the information from the abusive language.

Guang Xiang, Bin Fan, Ling Wang, Jason I. Hong, Carolyn P. Rose have proposed Detecting Offensive Tweets via Topical Feature Discovery over a Large-Scale Twitter Corpus [7]. A dataset of 860071 samples is collected and processed. They proposed a semi-supervised approach for detecting profanity-related offensive content in Twitter. This method exploits the lexical collocation of profane language via statistical topic modelling techniques and detects offensive tweets using highly expensive topical features.

Zhi Xu, Sencun Zhu has worked on Filtering Offensive Language in Online Communities using Grammatical Relations [8]. A dataset of 11,000 text comments is collected from the YouTube website. They proposed a new automatic sentence-level filtering approach which is able to semantically remove the offensive language by utilizing the grammatical relations among words.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi and Maurizio Tesconi have worked towards Hate me, hate me not: Hate Speech Detection on Facebook [9]. This proposed system uses two different algorithms in order to verify their classification performance on the task of hate speech recognition. The first is based on Support Vector Machine (SVM) when data is fed as an input it organizes into three main categories raw and lexical text features, morpho-syntactic and syntactic features, and lexicon features. Second method is Recurrent Neural Network named as Long Short-Term Memory (LSTM) here each input is represented by a 262-dimensional vector. The result shows the excellent

Table 1
Analysis

Paper	Methodology	Drawbacks
Offensive Language Detection Using Multi-Level Classification	This proposed system is taking advantage of a variety of statistical models and rule-based patterns, there is an auxiliary weighted pattern repository which improves accuracy by matching the text to its graded entries.	It does not consider the syntactical structure of the messages explicitly and could be equipped with some modules designed for subjectivity detection based on their lexicons (in this case it has to take it into account that the length of each message would be a limitation for the method).
Detecting Offensive Language in social media to Protect Adolescent Online Safety	In this proposed system, they incorporate a user's writing style, structure and specific cyberbullying content as features to predict the user's potentiality to send out offensive content. Results from experiments showed that their LSF framework performed significantly better than existing methods in offensive content detection. Meanwhile, the processing speed of LSF is approximately 10msec per sentence, suggesting the potential for effective deployment in social media.	It is a time taking process (Data preprocessing).
Towards Accurate Detection of Offensive Language in Online Communication in Arabic	They collected and labelled a large dataset of YouTube comments in Arabic which contains a broad range of both offensive and inoffensive comments. They used this dataset to train a Support Vector Machine classifier and experimented with combinations of word-level features, N-gram features and a variety of pre-processing techniques. They summarise the pre-processing steps and features that allow training a classifier which is more precise, with 90.05% accuracy, than classifiers reported by previous studies on Arabic text.	It is observed that data pre-processing with stemming can be leveraged to enhance the detection of offensive language in casual Arabic text used in social media platforms. In addition, the utilization of N-gram features improves the classifier's performance. However, the combination between stemming and N-gram features has negative effect on precision and recall in our experiments, thus it is conclude that it is not beneficial to use both stemming and N-gram features within the same machine learning process.
Abusive Language Detection in Online Conversations by Combining Content- and Graph- Based Features	Proposed fusion methods integrating content- and graph-based features. Experiments performed on raw chat logs show not only that the content of the messages, but also their dynamics within a conversation contain partially complementary information, allowing performance improvements on an abusive message classification task.	One limitation of this method is the computational time required to extract certain features. Another limitation is it uses small dataset.
Offensive Language Detection Explained	In this paper, they analyze and compare four explanation methods for different offensive language classifiers: an interpretable machine learning model (naive Bayes), a model-agnostic explanation method (LIME), a model-based explanation method (LRP), and a self-explanatory model (LSTM with an attention mechanism). They evaluate these approaches with regard to their explanatory power and their ability to point out which words are most relevant for a classifier's decision.	They find that the more complex models achieve better classification accuracy while also providing better explanations than the simpler models.

effectiveness of the two classification approaches.

3. Analysis

Table 1 shows the complete analysis of the existing systems in the chronological order.

4. Conclusion

Since the textual contents on online forum are highly unstructured, informal, and often misspelled, existing research on offensive language detection cannot accurately detect offensive content.

After the research conducted from the literature survey it has been revealed that the existing works use a lot of different approaches.

- *Multi-level Classification:* An automatic flame detection method which extracts features at different conceptual levels and applies multi-level classification for flame detection, the major drawback is it does not consider the syntactical structure of the messages.
- *Support Vector Machine:* This method is suited for only labelled data, and it is not easy to detect the offensive content in different languages, it is a time taking process because it uses variety of pre-processing techniques.
- *Combining Content-and Graph-Based Features:* The major drawback of this method is the computational time required to extract certain features and it is suitable for small datasets.
- *Filtering Method:* An automatic sentence-level filtering approach that is able to semantically remove the offensive language by utilizing the grammatical relations among words. The major drawback is the filtering might fail if offensive language cannot be detected before the filtering process.

Major drawbacks of the existing systems are:

- High computational time
- Datasets used are comparatively very small
- Less accuracy
- Detecting offensive content in only some particular languages i.e., existing systems are not suitable for all types of language.

References

[1] Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin "Offensive Language Detection Using Multi-Level Classification."

[2] Ying Chen, Sencun Zhu, Yilu Zhou, Heng Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety."

[3] Azalden Alakrota, Liam Murrayb, Nikola S. Nikolova "Toward Accurate Detection of Offensive Language in Online Communication in Arabic"

[4] Noe Cecillon, Vincent Labatut, Richard, Dufour and Georges Linares "Abusive Language Detection in Online Conversations by Combining Content- and Graph- Based Features"

[5] Julian Risch, Robin Ruff, Ralf Krestel, "Offensive Language Detection Explained."

[6] Altaf Mahmud, Kazi Zubair Ahmed, Mumit Khan "Detecting Flames and Insults in Text."

[7] Guang Xiang, Bin Fan, Ling Wang, Jason I. Hong, Carolyn P. Rose "Detecting Offensive Tweets via Topical Feature Discovery over a Large-Scale Twitter Corpus."

[8] Zhi Xu, Sencun Zhu "Filtering Offensive Language in Online Communities using Grammatical Relations."

[9] Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi1 and Maurizio Tesconi1 "Hate me, Hate me Not: Hate Speech Detection on Facebook.

[10] Georgios K. Pitsilis, Heri Ramampiaro and Helge Langseth "Detecting Offensive Language in Tweets using Deep Learning."

[11] Meredita Susanty, Sahrul, Ahmad Fauzan Rahman, Muhammad Dzaky Normansyah, Ade Irawan "Offensive Language Detection using Artificial Neural Network."

[12] Sravan Bodapati, Spandana Gella, Kasturi Bhattacharjee, Yaser Al-Onaizan "Neural Word Decomposition Models for Abusive Language Detection."

[13] Abdulaziz Saleh Ba Wazir, Hezerul Abdul Karim, Mohd Haris Lye Abdullah, Sarina Mansor, Nouar AlDahoul, Mohammad Faizal Ahmad Fauzi, John See, "Spectrogram-Based Classification of Spoken Foul Language Using Deep CNN."

[14] Abdulaziz Saleh Ba Wazir; Hezerul Abdul Karim; Nouar AlDahoul; Mohammad Faizal Ahmad Fauzi; Sarina Mansor; Mohd Haris Lye Abdullah; Hor Sui Lyn; Tabibah Zainab Zulkifli, "Spoken Malay Profanity Classification Using Convolutional Neural Network."

[15] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma, "Deep Learning for Hate Speech Detection in Tweets."

[16] Nemanja Djuric, Jing Zhou, Robin Morris, Mihaijo Grbovic, Vladan Radosavljevic, Narayan Bhamidipati, "Hate Speech Detection with Comment Embeddings."