An Effective Analysis of Anti Troll System using Artificial Intelligence

Aishwarya Gaikwad[1], Mrunmayee Patil[2], Sarang Patil[3], Mayura Rane[4], Dr. Shwetambari Chiwhane[5]

[1,2,3,4]B.E. Student, Dept. of Computer Engineering, NBN Sinhgad School of Engineering, Ambegoan, Pune- 411041, Maharashtra, India

[5]Professor, Dept. of Computer. Engineering, NBN Sinhgad School of Engineering, Ambegaon, Pune – 411041, Maharashtra, India

---***---

Abstract - *Online harassment has been on the rise ever since rampant boom in social media. Trolling is just another form of bullying that found its roots over the web. Certain anti-troll measures should be taken to deal with these issues and avoid promoting it further. Nowadays it has become a trend on social media to spew toxic hate. Some manual measures such as ignoring or blocking the trolls have been in use, but with the rise in the number of trolls, it needs more of an automated approach. Few social media platforms block trolls based on their set of troll words, however trolls resist these anti-troll systems by intentionally misspelling or other cunning methods. This paper discusses implementation of anti-troll using machine learning and artificial intelligence to provide a smarter troll detection system that adapts to current and updated trolling sense.*

Key Words: Sentiment Analysis, Artificial Intelligence, Anti Troll System, Twitter Sentiment Analysis

1. INTRODUCTION

Trolling on social websites has become very common activity nowadays. It is a huge issue in virtual world. As bullies have no restrictions from anyone they can easily get away after trolling a person. There is a need to create an application software to detect such kind of hazardous trolling and warn that bully so that he would think twice before doing such actions. Many companies have taken steps regarding trolling activities. There are very few software available which could detect foul words and simply block them but in troll detection system, our respective software system needs to have a clear understanding of sentences and clear language used by the troll. In this paper we are exploring various "Anti trolling systems" and different functionalities used in that system. This paper also suggests some machine learning algorithms and sentimental analysis used in anti-trolling systems. We are using Twitter social networking platform as an API for detecting our trolls using various methodologies. This paper also discusses about current functionalities and architecture used for anti-trolling systems and acquire a vision of trolling free internet.

2. RELATED WORK

2.1 Literature Survey on Sentimental analysis

The main objective of carrying out sentimental analysis is to detect trolls. During this process few important steps such as:

1. Task definition

2. Annotation guidelines

3. Data collection and annotation.

Under task definition few terms are used:

Repetitiveness- trolls send a large number of troll messages.

Destructiveness- troll messages express negative sentiments to sow discord

Deceptiveness: troll messages may be deceptive to achieve their objective of creating discord.

In annotation section, a troll message is a message that expresses negative sentiments, or irrational and offensive opinions. A troll is a person who posts a large number of such inflammatory messages.

In Data collection, we collect samples for annotation and then use the open source Hater News to score all users who had posted in at least 5 threads and at least 5 times in each of these threads. After ranking the users by their Hater News scores, we selected the 15 top-ranked users. For example, if he has been annotated as troll in 4 out of 5 threads, he has a troll-score of 0.8.

2.2 Literature Survey on supervised machine learning for troll detection

Using supervised learning, it is possible to link fake trolling account to the real account to see whether fake account is being used and check various features present in accounts and tweets using AI.

For doing this activity, we need to check the authorship of the account based on the comments posted by users. After doing this real cyberbullying techniques are applied on comments collected.

The authorship identification is done by analyzing on the group of profiles which have some common attributes present in them.

This step has 3 major steps which are:

- Selecting different profiles
- Collecting profiles data and tweets
- Features

Selecting different profiles comprises of gathering of similar profiles which were obtained in authorship identification technique. Profiles which have no similar content or there is no relation among them are ignored. Second step is data collection, it is important to notice the limitations imposed by Twitter API. The number of requests should not exceed 350 per hour as it limits the performance. Java-based collecting method are used for selected profiles, the user ID and the timeline tweets ,until having at least 100 genuine tweets with all abbreviations and slang.

There are few features which are used are the tweet, time of publication, language and twitter client

The first feature, the tweet, is the text published by the user, which gives us the possibility to determine a writing style, very differentiating for every user. The language and geoposition help in filtering and determining the authorship because users have certain behaviors which can be extrapolated analyzing these features. In spite of the possibility that users have many devices which they use to tweet, most of the times they use their own preferred client, which provides scope for one more filtering mechanism.

2.3 Literature Survey on Real time sentimental analysis

In real time sentimental analysis tweets which are generated online are collected and classified .In this process there are steps which are incorporated for doing this task. In the first step the tweets are collected which is the form of hashtags, comments etc. There is a limit of 5 to 1000 tweets which are streaming live. In the second phase the tweets are tokenized that means the sentence is broken down and segregated into phrases, symbols ,keywords, nouns ,adjectives Stemming is performed to reduce words to unify across the documents and make it easier to classify similar words under a category. Stemming is also applied so to reduce sentence without changing the actual meaning of the sentence.

Algorithm called bag of words is used to check occurrence of words so it becomes easy to classify. If the words in the tweet match the words in the positive Bag of words then it is classified as positive tweet, similarly if the word in the tweet matches the words in the negative Bag of words, then it is classified as negative tweet.

Lemmatization is used for finding dictionary meaning of the tokens obtained from tokenization. Last phase of this process is to visualize the output

Fig2.1 shows the flow of how the framework built works starting from taking input from user, till displaying output.

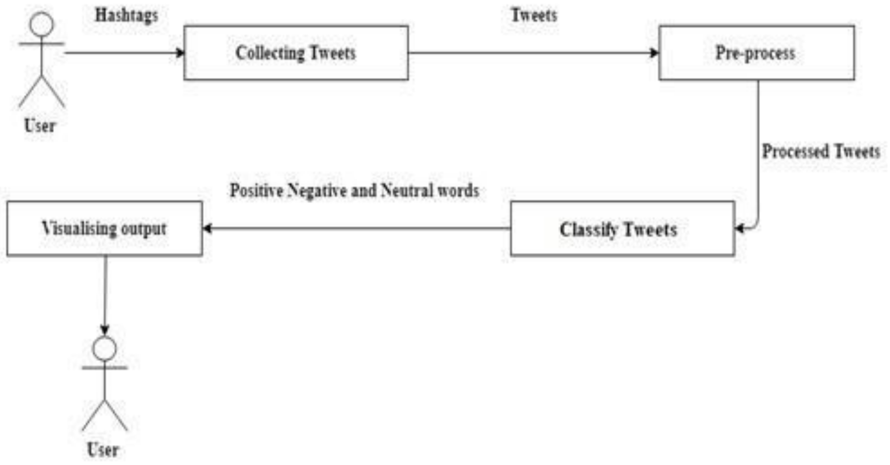

FIGURE 2.1: **Framework of sentiment analysis**

2.4 Literature Survey on Limitation of troll detection systems and AI/ML anti-trolling solution

Social network has become an open platform for everyone to write anything voluntary that further leads to cyberbullying activities. There are social websites like Facebook, Twitter and Google who have created their own basic systems for detecting foul and troll words. Facebook has its own Child Exploitation and Online Protection Centre (CEOP) they have created ClickCOEP button which are to detect and identify fake user accounts by keeping track of their IP addresses. Cyber bullying has taken wide place over the internet which causes Google and twitter to take a step forward and create software for online harassment. There is a company which has created the SMC4 (Social media C4) software which is first anti trolling software. This system detects trolls using various methodologies and algorithms. Then there is "Perspective API "which has been created by Google. It is Artificial Intelligent based system which uses conversational AI which could detect the harmful words and determine the toxicity level. Toxic words are removed out and conversational output is generated by the system. There are following steps for determining the toxicity level of the words

- Take input from user
- Dividing sentences into words for further analysis
- Remove the fillers
- Check the dataset for toxic words
- If the word is found define its toxicity
- Show output

We can find toxic words using lemmatization method which determines there toxicity level using sentimental analysis. There are misspelled words which can be reconstructed by other consulting words from the database (stemming).

2.5 Literature Survey on Troll Vulnerability in Online Social Networks

In this paper we mainly focus on different approaches of finding trolls, instead of detecting troll we focus on possible troll causing targets and detect it before getting trolled. We find out targets which are possible to be attracted by trolls using vulnerability of that post. To find out vulnerability of a post there is a define troll vulnerability rank (TVRank) metric. There is a model of troll vulnerability in which directed graph is used G = (v,e) where your nodes indicates posts and there is an edge from v post to u and v is a reply to u.

The vulnerability rank of the post should increase with the possibility towards trolling. The troll vulnerability results consist of accuracy model of precision and recall values under the ROC curve values of our graph.

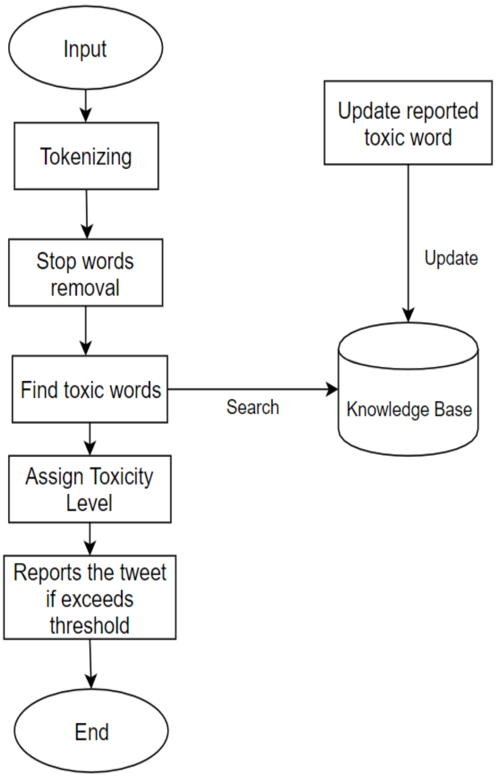

Figure 2.5.1 Working flow of Anti-troll system

Analysis of Literature Work:

Components	**Author**	**Methodology**
Sentimental analysis	Chun Wei Seah	Sentimental analysis forms a basic platform for finding trolls
Supervised machine learning for troll detection	Patxi Galan-Garcia	Identification of fake and real twitter profiles through supervised machine learning.
Real time sentimental analysis	Prakruthi V Sindhu D	Troll detection is done online. Collecting, pre-processing ,classifying and visualization is done
Limitations of troll detection systems and AI/ML anti-trolling solution	Ms. Swati Mali	Contextual analysis of words and determine toxicity level to it and removed out troll words
Troll Vulnerability in Social Networks	Paraskevas Tsantarliotis	It represents how trolls can be detected using directed graph technique based on vulnerability of a post

3. CONCLUSIONS

Based on this literature survey, presents an analysis of different techniques used in the process of troll detection. The paper discusses various previous state-of-art works such as Sentimental Analysis, Supervised ML for Troll Detection, Real time sentimental analysis, Limitations of troll detection system and AI/ML, Troll vulnerability in online social networks.

The first three papers are about the different methodologies being used for troll detections with sentimental analysis being predominant as it is an important part to detect the gist of the test using machine learning. The fourth paper discusses what kind of restrictions would be faced when using sentimental analysis for troll detection and how to overcome them. It brings to light various techniques that trolls use to surpass the troll detector.

REFERENCES

1) Troll-Detection Systems Limitations of Troll Detection Systems and AI/ML Anti-Trolling Solution / Ushma Bhatt / Divya Iyyani / Keshin Jani/Swati Mali / 2018IEEEhttps://ieeexplore.ieee.org/document/8529342.

2) Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying/Patxi Galán-García/José Gaviria de la Puerta/Carlos Laorden Gómez/Igor Santos/Pablo García Bringas/ 2016IEEEhttps://ieeexplore.ieee.org/document/8227084

3) Troll detection by domain-adapting sentiment analysis/ Chun Wei Seah / Hai Leong Chieu / Kian Ming A. Chai / Loo-Nin Teow / Lee Wei Yeong/ 2015IEEEhttps://ieeexplore.ieee.org/document/7266641

4) Real Time Sentiment Analysis Of Twitter Posts Prakruthi V /Sindhu D / Dr S Anupama Kumar/ 2018IEEE https://ieeexplore.ieee.org/document/8768774

5) Sentiment Analysis of Twitter Data /Sahar A. El Rahman/ Feddah Alhumaidi AlOtaibi / Wejdan Abdullah AlShehri/2018IEEEhttps://ieeexplore.ieee.org/document/8716464

6) A Hybrid Approach for Detecting Automated Spammers in Twitter /Mohd Fazil ; Muhammad Abulaish/2018IEEEhttps://ieeexplore.ieee.org/document/8335803

7) B. Pang and L. Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,” in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 115–124. [Online]. Available: http://dx.doi.org/10.3115/1219840.1219855

8) C. Laorden, B. Sanz, G. Alvarez and P. G. Bringas. A threat model approach to threats and vulnerabilities in online social networks. In Computational Intelligence in Security for Information Systems 2010, Vol. 85 of Advances in Intelligent and Soft Computing, pp. 135–142, 2010.

9) Zhao, Y. (2016). Twitter Data Analysis with R – Text Mining and Social Network Analysis. [online] University of Canberra, p.40. Available at:https://paulvanderlaken.files.wordpress.com/2017/08/rdataminingslides-twitter-analysis.pdf [Accessed 7 Feb. 2018].