First International Conference on Computer Engineering
International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
Volume 5 Issue 8, November-December-2020

Troll Detection and Anti-Trolling Solution using Artificial Intelligence or Machine Learning

Saloni Dangre[1], Shubham Sharma[1], Swati Balyan[1] , Tanisha Jaiswal[1] , Dr. Pankaj Agarkar[2], Prof. Pooja Shinde[3]
[1]Department of Computer Engineering, Dr. D.Y.Patil School of Engineering, Lohegaon, Maharashtra India
[2]Head of Department , Department of Computer Engineering, Dr. D.Y.Patil School of Engineering, Lohegaon, Maharashtra India
[3]Professor, Department of Computer Engineering, Dr. D.Y.Patil School of Engineering, Lohegaon, Maharashtra India

ABSTRACT

With the increase in usage of social media platforms, bullying and trolling has burgeoned proportionately. The sole reason for this is that there is no surveilling authority on these platforms. To add to that, anonymity protects the identity of these bullies. Anyone from kids to teenagers to adults can fall prey to trolling. This paper focuses on using AI/ML algorithms to invigilate and report such bullies and further take actions depending on the severity of the threat imposed by them. We will be introducing lexical, aggression, syntactic and sentiment analyzers to examine a tweet and determine if it was meant to be a troll or not. The output of these analyzers will be then fed to classifier algorithms such as Naive Bayes algorithm, K-mean, to segregate these tweets based on their toxicity rating.

Keywords : Social Media; Antisocial Behaviour; Troll Detection

I. INTRODUCTION

In recent years social media has been adopted in various countries by the general publicand also by companies. Additionally, “being social”, in contrast to “being a troll”, has been shown to be vital for the standard of human interaction within the digital sphere; this attitude is often assessed in different ways. A troll remains private with an antisocial behavior that incites other users acting within an equivalent social network. In particular, a troll often uses an aggressive offensive language and has the aim to hamper the normal evolution of a web discussion and possibly to interrupt it. Only recently has it been possible to pay proper attention to the present problem, in order that many renowned press bodies and magazines have begun to address the difficulty and to write down articles both on the overall description of the phenomenon and on particular events that have caused a stir, favored by the increasing occurrence of behaviorsjust like the one described above. This type of behavior isn’t fully characterized and, up to now, it’s been difficult to seek an accurate description for the word “troll”, since the act of trolling is strongly subjective. The shortage of an agreed-on definition for the term “troll” has resulted in poor comprehension and in low interest for the research community. The necessity for handlingthis problem has therefore emerged over time, alongsidestudies conducted by several universities and research centers. After removing applications which aren't strictly associatedwith the most topics taken into consideration (social sciences, computing and engineering), Scopus, as of 4 February 2020, lists 636 papers having the term "troll" within the title or abstract or as a keyword, when limiting the search to those three subject areas, 401 of which are related with the 2 latter topics, and 192 only to "computer science". Adding the keyword "detection" brings the entire right down to 51 papers, whose distribution in time shows a transparent increment after 2015.

For many people round the world social media sites are an integrated part of their lifestyle. There are many different social media sites supporting a good range of practices and interests. Social networks like Facebook and Twitter have become a source for news and a platform for political and moral debate for tons of users. Stories with different degrees of truthfulness are spread and tiny source criticism is applied by regular people also as journalists. The act of spreading disinformation on social media has developed from being caused by bored youths to being commercialized by organizations and political blocks within the sort of troll farms. A troll farm is a corporation whose sole purpose is to affect popular opinion with the means of social media. A practical implementation of a system or a software which will identify troll farms might be utilized in order to prevent them and thus avoid the spread of disinformation. Such an implementation would be interesting to the politicians, media, social networks or organizations that are targeted since it might be used to clear their names.

II. PREVIOUS WORK

Since 2004, with Orkut, Facebook and other social networks, people started sharing their opinions online with none moderation. Differences in opinions led to verbal spats, insults and slander. Trolling has become rampant with nobody to see on what people say within the virtual world. This inspired us to make an Anti-trolling system which will shield users online.

With the pleas of the general public getting louder and louder, Internet giants like Twitter, Facebook, Google etc. have come up with some solutions.

A. Facebook

Facebook has been the recentfavorite platform of paedophiles, bullies and hackers for an extended time. Presently users manually delete abusive messages. However, to tackle the emerging trend of online-trolls Facebook announced that they're performing on systems that automatically identify and delete abusive remarks. Administrators are being trained to handle trolls and given new tools to curb jibes. Children can report bullying and dubious behavior instantly to authorities. Official figures indicate that children below the permissible age of 13 use Facebook. Jim Gamble, the chief executive of the Child Exploitation and Online Protection Centre (CEOP) has been working with Facebook to combat cyberbullying.

Users who interact more with strangers and whose friend requests aren't accepted by other users are tagged. They maintain comprehensive grey lists to stop suspects from signing up using fake accounts. The reporting process helps to spot accounts having an equivalent IP address and each one of the fake accounts is removed.

Facebook is currently testing three new safety features around the world:

- One notifies users if someone is impersonating their account
- It analyses account names and profile pictures to find matches
- Another tool is for reporting nude photos and one more for photo safety

B. Twitter

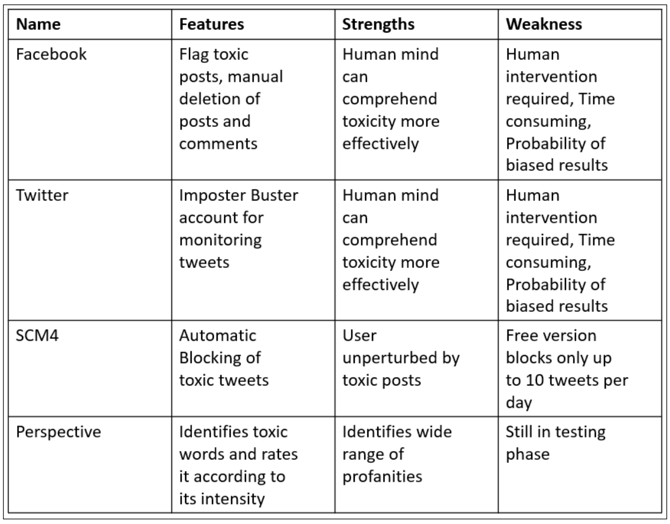
Twitter has often been a landmine of abusive language but they're trying to maneuver from being the Wild West to a more civilized society.

Twitter is cracking down harder against trolls, including temporarily barring accounts used to harass other users. During a blog post, Twitter's vice chairman of engineering, Ed Ho announced more safety measures to prevent abuse on its platform. "Making Twitter a safer place is our primary focus and that we are now moving with more urgency than ever," Ho said in a post.One among the methods includes using the company's internal algorithms to spot problematic accounts and limiting certain account functions for a selected period of time. Twitter is additionally hospitable to further action if the harassment continued. Other anti-trolling tools include new filters to let users see what type of content they need to look at from certain accounts. They also allow people to "mute" tweets supported keywords, phrases or entire conversations. The announcement follows a series of measures that Twitter has undertaken to curb abusive behavior on its platform. The company said that it might get obviate potentially harassing tweets from feeds and searches. It also has blocked people who repeatedly swore at verified accounts.

Twitter created an automated Twitter account, Imposter Buster, which is programmed with an updated database of impersonator accounts, and every time one among them vitriolic tweets, he automatically replies and exposes them with pre-programmed evidence. But this wasn't a really foolproof method as trollers could create multiple accounts to evade the anti-trolling account.

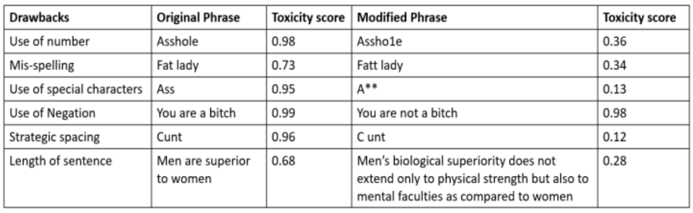
III. LIMITATIONSOFCURRENTSOLUTIONFORON LINE SOCIALNETWORK:

Table 1- Comparison of existing anti-trolling systems



Name	Features	Strengths	Weakness
Facebook	Flag toxic posts, manual deletion of posts and comments	Human mind can comprehend toxicity more effectively	Human intervention required, Time consuming, Probability of biased results
Twitter	Imposter Buster account for monitoring tweets	Human mind can comprehend toxicity more effectively	Human intervention required, Time consuming, Probability of biased results
SCM4	Automatic Blocking of toxic tweets	User unperturbed by toxic posts	Free version blocks only up to 10 tweets per day
Perspective	Identifies toxic words and rates it according to its intensity	Identifies wide range of profanities	Still in testing phase

Table 2- Issues in perspective



Drawbacks	Original Phrase	Toxicity score	Modified Phrase	Toxicity score
Use of number	Asshole	0.98	Assho1e	0.36
Mis-spelling	Fat lady	0.73	Fatt lady	0.34
Use of special characters	Ass	0.95	A**	0.13
Use of Negation	You are a bitch	0.99	You are not a bitch	0.98
Strategic spacing	Cunt	0.96	C unt	0.12
Length of sentence	Men are superior to women	0.68	Men's biological superiority does not extend only to physical strength but also to mental faculties as compared to women	0.28

IV. PROPOSEDWORK

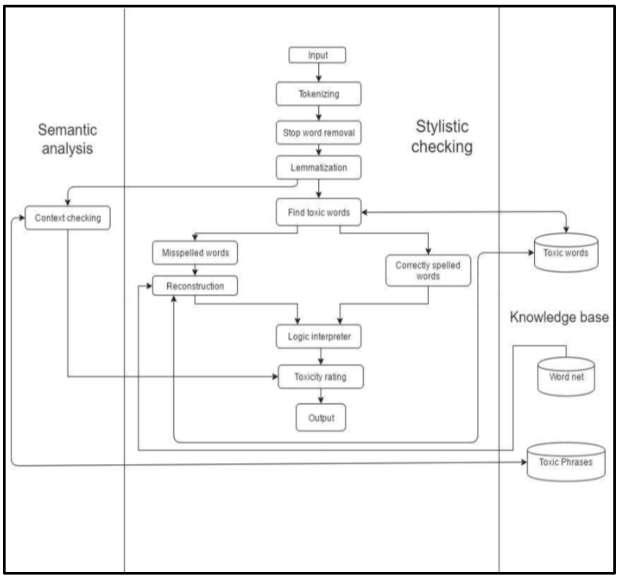


Fig. 1- Proposed Architecture

A. **Semantic Analysis:**

In this layer, the contextual meaning of the sentence is going to be analyzed.

Context checking: The precise meaning of the sentence can'tbe always understood by the literal meaning of the words utilized in the sentence. Hence during this part, the contextual meaning is taken into account.

B. **Stylistic Checking:**

Input: The tweets and other inputs are going to be accepted here in text format.

Tokenizing: The given sentence can't be easily understood by considering the whole sentence in one go. Hence, the sentence is weakened into the little part, i.e. one word per part referred to as a token. This manner helps in better understanding of the sentence.

Stop Word Removal: The words which don't contribute within the increase of toxicity of the sentence are mentioned as stop words. Such words (e.g. the, and, or) are deleted from the sentence during this step.

Lemmatization: The basic form of a word or its dictionary form is named lemma. Hence, during this part of the method, the basic form of the word is going to be returned, which can help in removing the inflectional endings and can make the method easier.

Find toxic Words: As soon as the sentence is successfully converted into tokens, the words are checked within the database for a match. (words here mean the foul, vulgar or gross language).

Misspelled Words Recognition: The words which are matched with the toxic words within the database are then forwarded to the logic interpreter after completion of indispensable reconstruction.

Correctly Spelled Words: The words which don't contain any toxicity are directly forwarded to the logic interpreted.

Logic Interpreter: After being recognized from the list and suitable reconstruction, the toxic word is passed on to the second last stage of processing where it's checked whether 'not' is employed before a toxic word. If used, as a result the comparator has got to minimize the toxicity level.

Toxicity Rating: Supported the severity of the toxicity of a word, a rating is going to be provided.

Output: The toxicity rating of the sentence as whole are going be displayed as the output.

C. **Knowledge Base:**

This domain comprises all the databases that are required for running the system. It embodies the word net, toxic words, their ratings and toxic phrases which will be appended during semantic analysis.

V. CONCLUSION

This article has discussed the problems created by the presence of trolls in social media contexts and has presented the main approaches to their detection. In conclusion, this project has been more or less a success. In the search for trolls in the Twitter network, we managed to find our own fabricated troll farm using only the daily activity and habits of the Twitter users. We will be using the K-mean and Naive Bayes algorithm. The results showed us that both algorithms could work and could be used in the search for trolls in a social network. The K-mean and Naive Bayes algorithm is used to segregate these inputs based on their toxicity rating. The drawbacks of the current system are solved in the proposed system. The problems of spacing, special characters, negation and other drawbacks are resolved. The stylistic drawbacks as well as the contextual drawbacks of the current system are solved with the use of different analyzers as mentioned above. In this paper, we took a different approach focusing on trolling vulnerability, negative comments and algorithms to detect trolling behavior. So, by working on troll vulnerabilities and their causes we can take proactive measures to stop or minimize troll or bullying nature.

VI. REFERENCES

[1]. Zannettou, S.; Sirivianos, M.; Caulfield, T.; Stringhini, G.; De Cristofaro, E.; Blackburn, J. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In Proceedings of the Web Conference 2019—Companion of the World Wide Web Conference, WWW 2019, San Francisco, CA, USA, 13–17 May 2019; pp. 218–226.

[2]. Badawy, A.; Lerman, K.; Ferrara, E. Who falls for online political manipulation? In Proceedings of the Web Conference 2019—Companion of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 162–168.

[3]. Fornacciari, P.; Mordonini, M.; Poggi, A.; Sani, L.; Tomaiuolo, M. A holistic system for troll detection on Twitter. Comput. Hum. Behav. 2018, 89, 258–268.

[4]. Donath, J.S. Identity and deception in the virtual community. In Communities in Cyberspace; Routledge: Abingdon-on-Thames, UK, 2002; pp. 37–68.

[5]. Chun, S.A.; Holowczak, R.; Dharan, K.N.; Wang, R.; Basu, S.; Geller, J. Detecting political bias trolls in Twitter data. In Proceedings of the 15th International Conference on Web Information Systems and Technologies, WEBIST 2019, Vienna, Austria, 18–20 September 2019; pp. 334–342.

[6]. "https://perspectiveapi.com/", last retrieved on 10th January 2017.

[7]. "https://jigsaw.google.com/projects/#conversation-ai", last retrieved on 14th March 2017.

[8]. "http://www.telegraph.co.uk/technology/internetsecurity/10283665/Worlds-first-anti-trolling-software-launchedin-UK.html", last retrieved on 10th October 2017.

[9]. "https://www.networkworld.com/article/2225302/opensourcesubnet/introducing-the-world-s-first--anti-trolling-software- .html" ,last retrieved on 10th October 2017.

[10]. "http://www.wired.co.uk/article/twitter-tools-harassment", last retrieved on 10th October 2017.

[11]. "http://www.dailymail.co.uk/sciencetech/article-3506068/Istrolling-Facebook-working-troll-

hunter-tool-lets-catchimpersonators-social-network.html”,last retrieved on 10th October 2017.

[12]. “https://thinkprogress.org/microsofts-lovable-teen-chatbotturned-racist-troll-proves-how-badly-silicon-valley-needsdiversity-1648e7020ea6#.dwoymxuwj”,last retrieved on 21st March 2017.

[13]. “http://www.tabletmag.com/scroll/219117/we-built-a-bot-thattrolls-twitters-worst-anti-semitic-trolls”, last retrieved on 21st March 2017.

[14]. “World’s first anti-trolling software' launched in UK”, “http://www.telegraph.co.uk/technology/internetsecurity/10283665/Worlds-first-anti-trolling-software-launchedin-UK.html”, last retrieved on 20th March 2017.

[15]. Phillips, Whitney; Publication Information: Cambridge, Massachusetts: The MIT Press. 2015 This Is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture.