Public Shaming Analyzer using Random Forest Classifier

Romin Katre
Dept. of Information Technology
Vidyavardhini's College of Engineering & Technology
Mumbai, India
romin.182094105@vcet.edu.in

Chirag Narkar
Dept. of Information Technology
Vidyavardhini's College of Engineering & Technology
Mumbai, India
chiragnarkar2507@gmail.com

Harsh Kore
Dept. of Information Technology
Vidyavardhini's College of Engineering & Technology
Mumbai, India
harshbkore@gmail.com

Abstract - Online social networks and public forums such as Twitter, Facebook, and Instagram have seen an increase in public shaming in recent years. This behavior can have a devastating impact on the victim's personal, political, and financial life. Despite the acknowledged negative effects, little has been done to address this problem on social media platforms, often due to the sheer volume and variety of comments, making it impractical for human moderators to handle. This study aims to automate the task of detecting public shaming on social media from the perspective of victims and identify common factors such as the shaming event and the shamers. The results show that the majority of users who participate in a shaming event contribute to victim-shaming, and the shamers often see a faster increase in their follower counts compared to non-shamers on social media. Finally, an online application has been developed to block and mute shamers attacking a victim, based on the categorization and type of shaming tweets/comments.

Keywords—Public Shaming, Social-Media, ShamersVictims, Random Forst Classifier.

I. INTRODUCTION

Public shaming is a form of social punishment that involves publicly calling out and criticizing individuals or groups for perceived wrongdoings or nonconformity to social norms. With the rise of social media platforms, public shaming has become more prevalent, and the impact can be devastating, often leading to social, psychological, and economic harm to the victim. While public shaming can serve as a form of social control, it can also be misused and abused, leading to negative consequences. As such, it is important to examine the effects of public shaming and explore ways to mitigate its harmful effects while preserving the potential benefits of holding individuals accountable for their actions[1].Insults and shaming, whether in person or on public platforms, are a universal experience for all individuals. These types of behavior can include abuse and discrimination based on various characteristics such as race, gender, sexual orientation, religion, and political beliefs. With the rapid growth of global connectivity, the number of instances of hate speech, shaming, and cyber bullying has increased significantly. Examples of such behavior include statements like "Women are scumbags" and "You don't deserve to live." This rise in such behavior highlights the alarming insensitivity of humankind. To address this issue, there is a need to detect online shaming and automate the process to mitigate its harmful effects. This paper contributes to this field by examining various aspects of shaming, such as the location of the comments, visual content, periodicity of postings, the targeted victim, and the community of offenders. The study presents a method for detecting shaming on social media platforms and taking automated actions against the user responsible for such behavior. Every individual is targeted online whether they have done something wrong or not. Even its normal human tendencies to have differences of opinion on a topic but are cursed, abused on the difference of opinion. One can hide their identity on various public platforms and can harass them to take any drastic step. One can also influence an opinion that may lead to making wrong decisions sometimes. Moreover, for a limited number of comments, you can personally delete/ block the particular comment. But comment sections for public figures have a multitudinous number of comments and dealing with it manually is something that we can say is a tedious job[2]. The paper's purpose is to learn how to make such a tedious job easy by using various Machine Learning algorithms by constructing a public shaming analyzer machine learning tool that gives you the list of comments which will classify is it abusive/ shaming or not. A public shaming analyzer can be a useful tool for understanding the impact of public shaming on individuals and society as a whole. Public shaming refers to the practice of publicly calling out and criticizing individuals or groups for their behavior or actions, often through social

media or other online platforms. One of the main benefits of a public shaming analyzer is that it can help to shed light on the effects of public shaming on mental health and well-being. Studies have shown that public shaming can lead to feelings of humiliation, anxiety, depression, and even suicidal thoughts. By analyzing patterns in public shaming behavior and the responses of those who are shamed, researchers and mental health professionals can gain a better understanding of how to prevent or mitigate the negative effects of public shaming. Additionally, a public shaming analyzer can help to identify cases of online harassment and abuse, which are often a byproduct of public shaming. By tracking patterns in the language and behavior of those who engage in public shaming, researchers and law enforcement officials can better understand the motivations and tactics of online harassers, and develop strategies for preventing and addressing these harmful behaviors. Overall, a public shaming analyzer can be an important tool for promoting empathy and understanding in online communities, and for working towards a more compassionate and supportive society. By encouraging more thoughtful and respectful communication online, we can help to reduce the harm caused by public shaming and other forms of online abuse. A random forest classifier can be a useful tool for analyzing instances of public shaming on social media platforms. Random forest classifiers are machine learning algorithms that can be used for both classification and regression tasks. They work by building a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

To train a random forest classifier for public shaming analysis, a dataset of labeled instances of shaming and non-shaming comments on social media platforms can be used. The dataset can be preprocessed to extract relevant features, such as the language used, sentiment, the presence of certain keywords, and user information (e.g., number of followers, past behavior). These features can be used to train the random forest classifier, which can then be used to predict whether new instances of social media comments represent instances of public shaming. The output of the random forest classifier[3] can then be used to develop automated tools for identifying and mitigating instances of public shaming. For example, comments classified as instances of public shaming could be flagged for further review by human moderators or automatically hidden from view to reduce their impact on the targeted individual. Another study [4]examined the relationship between online social networking and the experience of cyber-bullying. Online social networking has been linked to the experience of cyber bullying, as social media platforms provide a space where individuals can easily communicate with each other, both positively and negatively. Cyber bullying is defined as the use of technology to harass, embarrass, or threaten another person. This can take many forms, such as spreading rumors or posting hurtful comments, images or videos online. Social media platforms provide individuals with the ability to communicate with each other in real-time, with the potential to reach a large audience[4], [5]. This can make cyber bullying more pervasive, as hurtful messages or comments can spread quickly and easily. Additionally, the anonymity that social media platforms provide can make it easier for individuals to engage in cyber bullying without fear of consequences. Research has found that individuals who use social media are more likely to experience cyber bullying than those who do not use social media. Additionally, individuals who spend more time on social media platforms are more likely to experience cyber bullying than those who spend less time on these platforms. The age group most affected by cyber bullying is young people, especially teenagers, who are more likely to use social media platforms. Overall, the relationship between online social networking and cyber bullying is complex, as social media can be used for both positive and negative interactions[6], [7]. However, it is important for individuals to be aware of the potential risks associated with social media use and take steps to protect themselves, such as blocking and reporting cyberbullies and limiting their online exposure[8].In this research the first extraction of the comments from a particular account will be done after going through the comment section of all the posts of a particular account, by using a machine trained by one of the algorithms of machine learning it will be classified as shaming or not.

II. METHODOLOGY

In this section, we present our procedure for making a public shaming analyzer with the highest accuracy possible. The procedure is divided into smaller steps to attain maximum accuracy, success, and is free of bugs. The main objective of this project is to analyze social media comments. The entire project is divided into three major steps as shown in figure 1:

- Scraping of the comments
- Pre-processing of data acquired
- Training and Testing of the Machine Learning Model.

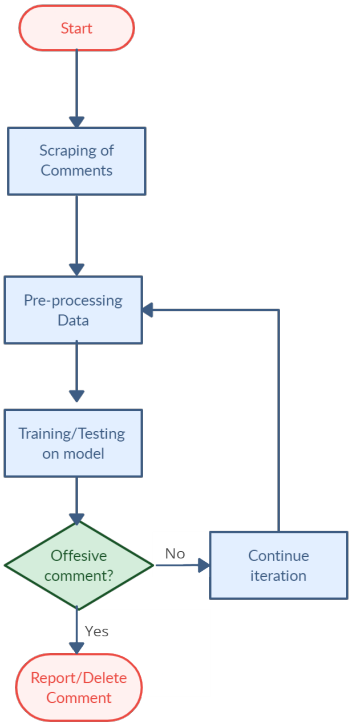

Fig. 1. Procedure Flow

A. *Scraping of the comments*

The first stage of our project involves gathering real-time comments on a specific account. This was accomplished by utilizing selenium to scrape the comments in real-time. Selenium is an automation tool commonly used for testing web applications. In order to gain control of the Chrome web browser, which is necessary for the algorithm to visit the URL of each post and extract comments, a chrome driver is also required.

B. *Preprocessing of the data acquired*

Preprocessing of data is a crucial step in machine learning that involves preparing raw data for analysis. This step typically involves several processes, such as cleaning, normalization, transformation. Data cleaning involves removing any irrelevant, incomplete, or erroneous data, and handling missing values. Normalization involves scaling the data to ensure that each feature is within the same range, making it easier to compare across features. Transformation involves converting the data into a suitable format for the machine learning algorithm to process.

C. Training *and testing of the Machine Learning model*

The next and most important step of the project is to train and test the model. For Training our model we have used a random forest algorithm. For training, there were many dataset on open source, combining it and manually annotations were done to train from every aspect. Dataset was split in an 80:20 ratio where 80% were used in training and 20% was used in testing.

III. Solution Approach

Solution approach is critical for solving complex problems effectively and efficiently. It involves careful planning, analysis, and execution to ensure that the problem is fully understood, and the most appropriate solution is implemented. This section provides a comprehensive overview of the work conducted in the three distinct steps of the project.

A. *Scraping of the comments*

For Automating our web app on browsers selenium is used. Selenium has different functions. For their search, we have used a selenium web driver. Scraping comments with Selenium and Selenium WebDriver involves using automation tools to extract comments from a web page or application. Selenium is a popular automation tool that is commonly used for testing web applications, while the Selenium WebDriver [9] is a web automation framework that enables interaction with web browsers. To scrape comments with Selenium and Selenium WebDriver, the first step is to identify the web page or application where the comments are located. Once identified, the Selenium WebDriver is used to access and control the web browser, enabling the program to navigate to the relevant pages and extract the comments. In the study [10] it was observed that Selenium and Selenium WebDriver provide a range of features for web scraping, including the ability to interact with web forms, handle JavaScript, and navigate through multiple pages. These tools can also be used to automate the process of scrolling through comments and loading additional pages as required. Overall, scraping comments with Selenium and Selenium WebDriver can be a powerful tool for data collection and analysis. However, it is important to ensure that the scraping process is done in compliance with relevant legal and ethical standards, and that the data is used responsibly and appropriately.

B. *Pre-processing, Training, and Testing of the model*

1) *Pre-processing*

The subsequent stage after acquiring the unprocessed comments from a social media account involves pre-processing, which is crucial since the data set used to train the model often contains irrelevant or outlier elements that do not contribute to determining the comment's sentiment. For instance, in the comment "How is life going in Seattle,"

the words "is," "the," and "in" do not carry much significance and can be disregarded. The elimination of such "stop words" is elaborated further with other pre-processing methods.

2) Stop-words

These are the words that are the most commonly used in a language. In English the stop words can be “the”, “us”, “our”, “in” etc. All these types of words need to be removed as we are working on text classification, these words don’t give us any information about the text and hence can’t be given to a model for training and restricting the unwanted data from our corpus.

3) Stemming

This is the process of producing the variants of a root word or reducing a word to its root word. Considering an example, the words “Likes”, “Liked”, “Likely” and “Liking” can be reduced to their root words which are “Like”. But this cannot be applied in every case as it is prone to being erroneous hence lemmatization is also employed.

4) Lemmatization

This approach is similar to stemming i.e., reducing the words to their root words, but lemmatization is much more widely applicable as while converting the words to their roots the context of the original words is also considered and taken care of. Let’s say we have the word “Caring” if stemming is used the root word comes out to be “Car” but if the lemmatization approach is used the output is changed to “care” and it is evident from here that the context of the original words is preserved.

5) Tokenization

It is a process of splitting a sentence or paragraph into small units called tokens. This is an important step if we want to get the meaning of the sentence given, as the words present in the sentence gives us the meaning of the sentence rather than considering a whole sentence. For example, “Technology is Good” can be tokenized into [“Technology”, “is”, “Good”]. This helps us to determine the number of words in the sentence, it can also help us get information about the frequency of a particular word in the sentence.

6) Numeric and Special character removal

As we are doing text analysis, numeric values and special characters don’t play a major role in determining the meaning of the sentence hence has to be removed from the raw message.

7) Normalizing the slang

When working with social media comments, it is common to encounter slang language. However, models cannot be trained on slang words, so it is essential to convert them into their original forms to extract useful information that can inform decision-making in the future. For instance, the phrase "I luv u" should be converted to "I love you" to make sense. This requires a large dictionary of slang words and their corresponding original words. Here is a brief example of such a dictionary: slangdict = 'luv':'love', 'wud':'would', 'lyk':'like', 'wateva':'whatever', 'ttyl':'talk to you later', 'kul':'cool', 'fyn':'fine', 'omg':'oh my god!', 'fam':'family', 'bruh':'brother', 'cud':'could', 'fud':'food'.

8) Separating Hashtags

As we have comments as our inputs it is not surprising to have hashtags, we have to store hashtags separately as they generally hold weight in determining the context of a sentence as positive or negative.

C. Training and Testing

To train our model, we utilized the random forest algorithm. Several datasets were available on open-source platforms, and we combined them while manually annotating the data to ensure comprehensive training from all angles. We split the dataset in an 80:20 ratio, with 80% used for training and 20% reserved for testing.

IV. Results and Discussion

Random forest can be a useful algorithm for public shaming detection. With the increasing use of social media, instances of public shaming have become more common. Random forest can be used to analyze social media comments and detect instances of public shaming. By training the algorithm on datasets that include examples of public shaming, the model can learn to identify patterns and features associated with such behavior. One advantage of using random forest for public shaming detection is that it can handle a large number of features and variables. Social media comments can be complex, and it is essential to consider multiple factors when detecting instances of public shaming. Random forest can efficiently analyze a vast number of variables and identify the most relevant ones for detecting public shaming. Additionally, random forest is a flexible algorithm that can be adapted to different types of data. It can be used with both numerical and categorical variables, which makes it suitable for analyzing social media comments that often contain a mix of different types of data. Overall, random forest is a powerful tool that can be leveraged to detect instances of public shaming on social media platforms. It offers the potential to analyze large

volumes of data and identify patterns associated with public shaming behavior, which can help individuals and organizations take steps to prevent or address such behavior.

Fig. 2. Negative Words

The random forest algorithm uses bagging to train multiple decision trees, and the final output is determined by the majority vote of the individual tree predictions. Since each tree may provide different results, the one with the highest number of votes is selected and presented as the overall prediction. This approach enables the random forest algorithm to produce a single output based on the collective inputs from multiple decision trees. Key parameters that can affect the outcome include the node size, number of trees, and number of sampled features. A significant advantage of using random forests is their ability to minimize the risk of overfitting.

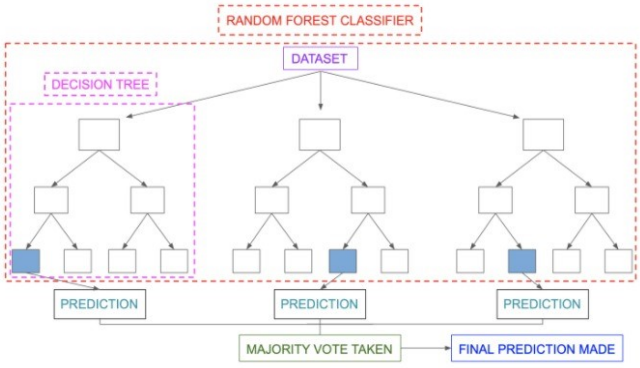

Fig. 4. Random Forest Architecture

Figure 5 provides the results from random forest using stemming and lemmatization. Stemming and lemmatization are text preprocessing techniques used to reduce words to their root form. Stemming involves removing the suffix of a word to obtain its stem, while lemmatization involves reducing a word to its base form, also known as a lemma. The accuracy of 0.94 suggests that the random forest classifier with stemming and lemmatization is performing very well on the task of identifying public shaming in text. This is a high level of accuracy and suggests that the model is a reliable tool for identifying public shaming in text. It is also observed from the figure that the F1 score is the harmonic mean of precision and recall, and provides a single metric to evaluate the overall performance of the model. An F1 score of 0.95 indicates that the model is performing well in both precision and recall, and is able to balance these metrics effectively. Figure 6 shows the results of random forest without stemming and lemmatization, it provides accuracy of 0.83, whereas F1 score is 0.84. As it is impossible to cover every emotion of a comment in this study, we have illustrated the results from figure 7-10 which provides actual and extracted comment from social media platforms such as Instagram and Facebook. Figure 11, provides a clear picture about the real time comment classified as hate speech. The ability to classify comments in real-time can help social media platforms quickly remove hate speech and prevent it from spreading, which can have a positive impact on the online community. Additionally, identifying hate speech in real-time can help to create a safer and more inclusive online environment for all users.

In this study random forest is utilized to analyse the public shaming data. Figure 2 and 3 illustrate the negative and positive words respectively in social media platforms.

Fig. 3. Normal Words

```
...  Random Forest Classifier Report
     Predicted     0     1
     Actual
     0          3221   166
     1           166  2122
                   precision    recall  f1-score   support

                0       0.95      0.95      0.95      3387
                1       0.93      0.93      0.93      2288

         accuracy                           0.94      5675
        macro avg       0.94      0.94      0.94      5675
     weighted avg       0.94      0.94      0.94      5675
```

Fig. 5. Random Forest report (with stemming and lemmatization)

```
Random Forest Classifier Report (Without Stemming and Lammetization)
Predicted     0     1
Actual
0          3215   172
1           168  2120
              precision    recall  f1-score   support

           0       0.80      0.85      0.84      3387
           1       0.82      0.82      0.81      2288

    accuracy                           0.83      5675
   macro avg       0.84      0.84      0.86      5675
weighted avg       0.84      0.84      0.86      5675
```

Fig. 6. Random Forest report (without stemming and lemmatization)

Fig. 7. Actual Comments(Instagram)

Fig. 8. Comment Extracted (Instagram)

Fig. 9. Actual Comments (Facebook)

Fig. 10. Comment Extracted (Facebook)

Fig. 11. Real-time comment classified as hate speech

V. Future Work

For future work, the present project solely focuses on analyzing comments. However, in the future, we could use deep learning methods to predict a user's personality based on their comment section, posted content, and search history. Furthermore, various methods could be employed to enhance the accuracy of the random forest method.

VI. Conclusion

Numerous recent studies have focused on text summarization. However, the use of text summarization in the context of cyber bullying requires more attention given the widespread growth of hate speech and cyber bullying on social media. Analyzing comments on a particular social media post can be useful in various scenarios, such as on a personal level for maintaining social platform integrity or for businesses to monitor their public image, which can be challenging for large businesses with extensive public interaction. In this study, we implemented a Random Forest algorithm for text analysis, and obtained a good result of 83% accuracy without Stemming and Lemmatization. However, the accuracy can be significantly improved by using stemming and lemmatization, which yielded a result of 94% accuracy. This demonstrates that Random Forest is an effective approach for text summarization and can be used for future work in this area. It is likely that automatic cyber bullying detection will become more efficient and useful than manual checks in the near future.

Acknowledgement

We would like to extend our sincere appreciation to Prof. Swati Verma, our internal guide, for providing invaluable assistance and support throughout the implementation of our idea. Additionally, we are grateful to our parents and friends, who helped us complete the paper within the specified timeframe. We are also thankful to our management faculty for their unwavering support. This paper is a result of the combined efforts and commitment of all our group members. The success and results of this endeavor were made possible by the guidance of many individuals, and we consider ourselves fortunate to have received such assistance.

References

[1] R. Basak, S. Sural, N. Ganguly, and S. K. Ghosh, "Online Public Shaming on Twitter: Detection, Analysis, and Mitigation," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 208–220, Apr. 2019, doi: 10.1109/TCSS.2019.2895734.

[2] P. Billingham and T. Parr, "Online Public Shaming: Virtues and Vices," *Journal of Social Philosophy*, vol. 51, no. 3, pp. 371–390, 2020, doi: 10.1111/josp.12308.

[3] S. S. Mohite, V. Attar, and S. Kalamkar, "Shaming tweets detection on Twitter using Machine learning Algorithms," in *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, Oct. 2022, pp. 1–6. doi: 10.1109/GCAT55367.2022.9972100.

[4] B. O'Dea and A. Campbell, "Online social networking and the experience of cyber-bullying," *Stud Health Technol Inform*, vol. 181, pp. 212–217, 2012.

[5] S. Abarna, J. I. Sheeba, S. Jayasrilakshmi, and S. P. Devaneyan, "Identification of cyber harassment and intention of target users on social media platforms," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105283, Oct. 2022, doi: 10.1016/j.engappai.2022.105283.

[6] S. Burke Winkelman, J. Oomen-Early, A. Walker, L. Chu, and A. Yick-Flanagan, "Exploring Cyber Harassment among Women Who Use Social Media," *Universal Journal of Public Health*, vol. 3, no. 5, pp. 194–201, Sep. 2015, doi: 10.13189/ujph.2015.030504.

[7] W. Akram, "A Study on Positive and Negative Effects of Social Media on Society," *International Journal of Computer Sciences and Engineering*, vol. 5, Mar. 2018, doi: 10.26438/ijcse/v5i10.351354.

[8] J. Amedie, "The Impact of Social Media on Society," *Pop Culture Intersections*, Sep. 2015, [Online]. Available: https://scholarcommons.scu.edu/engl_176/2

[9] S. Gojare, R. Joshi, and D. Gaigaware, "Analysis and Design of Selenium WebDriver Automation Testing Framework," *Procedia Computer Science*, vol. 50, pp. 341–346, Jan. 2015, doi: 10.1016/j.procs.2015.04.038.

[10] M. Bures and M. Filipsky, "SmartDriver: Extension of Selenium WebDriver to Create More Efficient Automated Tests," in *2016 6th International Conference on IT Convergence and Security (ICITCS)*, Sep. 2016, pp. 1–4. doi: 10.1109/ICITCS.2016.7740370.