Unmasking cyberbullying: Classifying comments and revealing communities

(CIS 600 | Social Media Data Mining Project Proposal)

04/03/2023

Team:

SN.	Name	SUID	Email
1	Shivakumar Suresh	376148007	shsuresh@syr.edu
2	Sanjay Kumar Thovinakere Srinivas	761896187	sthovina@syr.edu
3	Sujay Vishwanath Malghan	314885101	svishwan@syr.edu
4	Rakshitha Kandavara Jayarama	862972326	rkandava@syr.edu
5	Chaitra Sampathram	970565905	csampath@syr.edu
6	Bindushree Huruhihakkalu Ambikanath	375144255	bhuruhih@syr.edu
7	Deepika Nandan	927509909	dnandan@syr.edu
8	Riya Jomy Kannampuzha	303601108	rjk100@syr.edu

Abstract

Cyberbullying is a pervasive issue on social media platforms, adversely affecting the mental health and well-being of victims. We propose a comprehensive framework to unmask cyberbullying by classifying comments and revealing communities on social media. Leveraging techniques from previous works such as "Cyberbullying Detection: An Overview," "Cyberbullies in Twitter: A Focused Review," "Cyber-Bullying Detection: A Comparative Analysis of Twitter Data," "Cyber Bullying Detection Based on Twitter Dataset," and "Cyberbullying Detection Through Sentiment Analysis," we employ machine learning models to identify and categorize instances of cyberbullying. Our approach involves data collection, preprocessing, feature extraction, model selection, training, and evaluation. We experiment with different machine learning models such as Support Vector Machines, Naive Bayes, and Neural Networks to achieve optimal performance.

To reveal communities on social media platforms, we utilize network analysis techniques to perform community detection based on user interactions. We then apply our trained models to classify the comments of each user, allowing us to calculate the prevalence of cyberbullying in social media platforms and detect communities behind it. This process provides insights into the dynamics and relationships within the communities, enabling targeted interventions to mitigate the impact of cyberbullying.

Our results demonstrate the effectiveness of the proposed framework in accurately classifying instances of cyberbullying and uncovering communities with varying degrees of cyberbullying prevalence. Through visualizations of community behavior patterns, we gain a deeper understanding of the online environment and the factors that contribute to the presence of cyberbullying in different communities.

Our findings demonstrate the effectiveness of the proposed framework in accurately classifying cyberbullying instances and uncovering communities with varying degrees of cyberbullying prevalence. This project contributes to creating a safer online environment by providing insights into the behavior of different communities and identifying areas that may require intervention or support.

Contents

1. Introduction

Cyberbullying is bullying that occurs through the use of digital technology, such as cell phones, computers, and tablets. Cyberbullying can take place via SMS, text, and applications, as well as online in social media, forums, and gaming where people can see, engage in, or exchange content. Cyberbullying is defined as sending, uploading, or spreading nasty, harmful, false, or derogatory content about another person. It can include disclosing personal or private information about another person that causes embarrassment or humiliation. Several forms of cyberbullying are illegal or criminal in nature.

In the context of social media and network analysis, community refers to a group of users who share common interests, engage in similar activities, or interact with each other frequently. These users may be connected through friendships, followers, or other types of relationships on the platform. Communities in social media networks often exhibit a higher degree of interaction and communication among their members compared to users outside the community. Identifying such communities can help researchers and analysts better understand the dynamics, behavior patterns, and relationships within groups of users, which may be useful for targeted interventions or support, particularly when addressing issues such as cyberbullying.

The objective of this project is to create a classification model that can recognize various forms of toxic comments in online interactions and reveal communities, with the goal of preventing online bullying. The model will be trained on a labeled comment dataset to distinguish patterns and attributes associated with various sorts of toxicity, such as insults, threats, hate speech, and harassment. After training, the model may be used to examine fresh comments and categorize them into one or more toxicity categories.

Despite the advancements in cyberbullying detection, there is still a need to develop a comprehensive framework that not only classifies comments but also reveals communities on social media platforms based on their behavior. Uncovering communities with varying degrees of cyberbullying prevalence can provide valuable insights into the dynamics and relationships within these communities, enabling targeted interventions and support for those affected by cyberbullying.

The application of such a model has various possible advantages. It can assist in identifying and flagging potentially damaging comments prior to their posting, preventing them from causing harm. It can also help moderators review warned comments more efficiently, which saves time and resources. It can also help to create safer online environments by encouraging respectful and constructive conversations.

2. Application and its significance

A brief explanation of the application :

This project titled ***"Unmasking cyberbullying: Classifying comments and revealing communities"*** *aims* to build a machine learning model focused on classifying toxic comments and preventing online bullying. The dataset for this project will be extracted from Twitter using Twitter API and parsed using python libraries such as Tweepy and stored in a CSV file. This dataset will then be used to train a classification model that can detect different types of toxic comments, such as those related to identity-based hate, threats, insults, and obscenity. This model can also be used to analyze the type of toxic comments appearing online and help identify potential cases of online bullying or harassment. We use these data points to create a social network of size n and then identify the common users among them. This will help us to conclude if the comments were made by a specific community which can then be associated with their agenda and eventually be reported. Libraries and tools such as scikit-learn, Keras, and TensorFlow will be used for building the machine learning model.

Overall, this project aims to tackle the growing problem of online bullying and toxicity by providing a tool that can help identify and prevent toxic comments in real-time, thus creating a safer online environment for everyone.

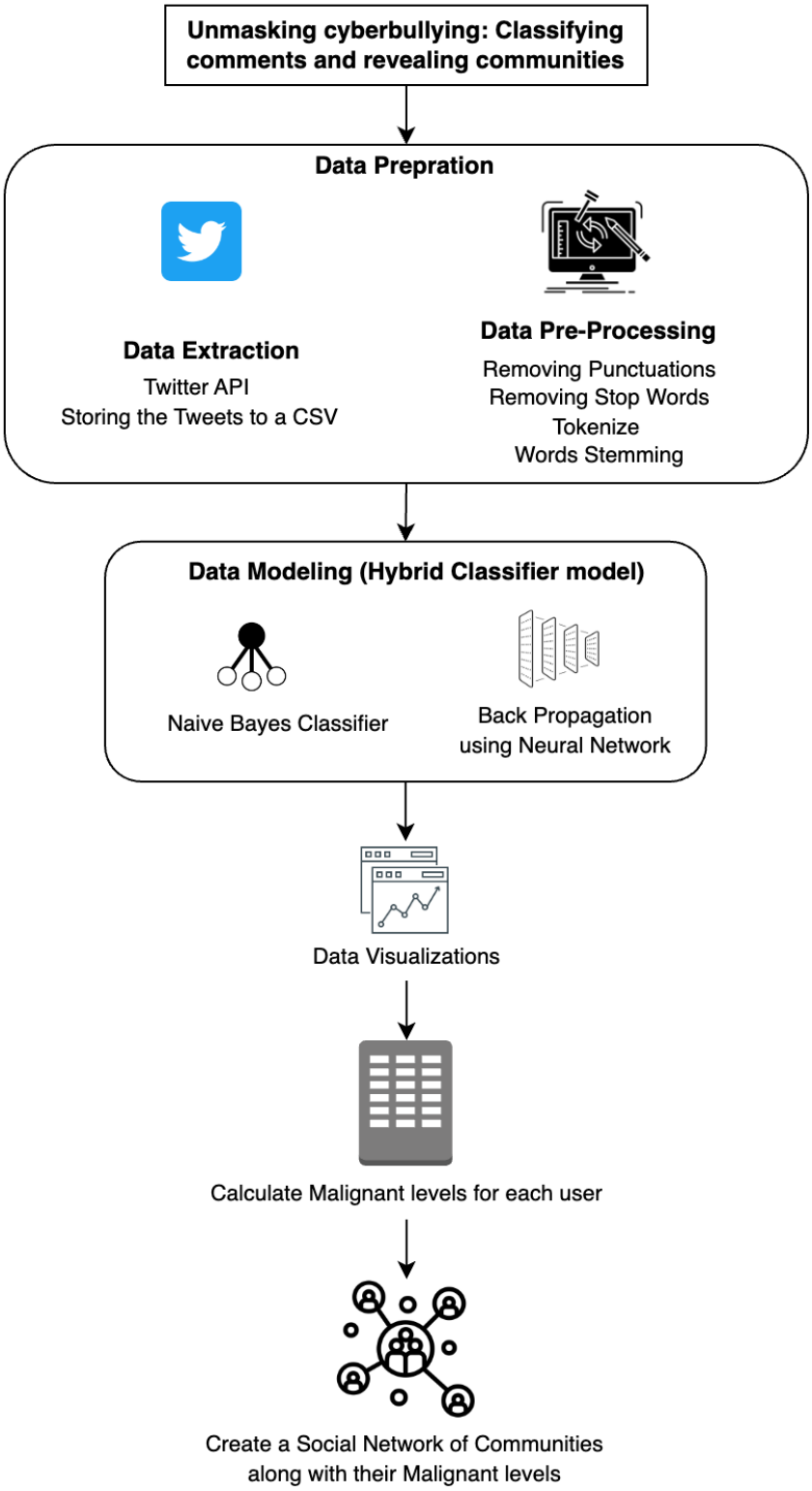
What is the significance of this application?

There are several compelling reasons why we believe that this project is worthwhile and has the potential to make a valuable contribution to the field of cyberbullying research:

- **Addressing a critical social issue:** Cyberbullying is a serious social issue that can have significant impacts on individuals and communities. By focusing on this issue, our project has the potential to make a meaningful contribution to the field of social media and to society as a whole.
- **Leveraging technology to understand and combat cyberbullying:** Our project involves using social media and machine learning techniques to analyze instances of cyberbullying. This approach has the potential to provide insights that are difficult to obtain through traditional research methods, and can help develop more effective strategies for preventing and mitigating cyberbullying.
- **Multidisciplinary nature**: This project is a multidisciplinary effort that brings together researchers from a variety of fields, including computer science, psychology, and sociology. This project can benefit from this interdisciplinary approach, as it can draw on insights and expertise from multiple fields to develop a more comprehensive understanding of social media and its impacts.

- **Potential for real-world impact:** Finally, our project has the potential to make a real-world impact by informing the development of interventions and strategies for addressing cyberbullying on social media. By identifying patterns and communities of cyberbullying, and by predicting its impact, this can help develop more effective interventions that can reduce the harm caused by cyberbullying.

3. Architecture diagram



4. Dataset exploration and feature extraction

4.1 Dataset Exploration for Classification :

The dataset for training and testing for our project will be the Toxic Comment Classification Challenge [https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge] obtained from the Kaggle platform.
The file train.csv, compressed in the zip folder, is the training set, which contains comments with their binary labels. The file train.csv has a total of 151203 samples of comments and labeled data.
The dataset consists of the following fields -

- id: An 8-digit integer value, to get the identity of the person who had written this comment
- comment_text: A multi-line text field which contains the unfiltered comment
- toxic: binary label which contains 0/1 (0 for no and 1 for yes)
- severe_toxic: binary label which contains 0/1
- obscene: binary label which contains 0/1
- threat: binary label which contains 0/1
- insult: binary label which contains 0/1
- identity_hate: binary label which contains 0/1

The comment_text field will be preprocessed and fed into different classifiers to predict whether it falls under one or more of the labels or outcome variables, namely toxic, severe_toxic, obscene, threat, insult, and identity_hate. The dataset is noticing the frequency of occurrence of multi-labeled data; if all the labels are 0, then the comment_text is classified as non-toxic. If any label has value 1, then the comment_text is classified as toxic comment.. Overall, 9790 samples are those that have at least one label, and 5957 samples have two or more labels.

4.2 Feature Extraction:

The track parameter used to collect the data is mainly the cuss words, as a result of which almost all of the collected data could be categorized as one of the malignant comment classes for the purpose of analyzing the predicted results.

The following steps are performed to achieve the results:

1. To set the authorization to the Twitter account.
2. To create the twitterStream.
3. Using the twitterStream to filter the data with the required words.
4. To extract the useful features from the twitter:
   a. tweetID

b. dateTime
c. tweet
d. screen_name
e. followers_count
f. source
g. country
h. country_code
i. full_name
j. name
k. place_type
l. quote_count
m.reply_count
n. retweet_count
o. favorite_count

5. To write the extracted data into the.csv file.

4.3 Data pre-processing:

Data preprocessing includes the steps we need to follow to transform or encode data so that it may be easily parsed by the machine. The main requirement for a model to be accurate and precise in predictions is that the algorithm should be able to easily interpret the data's features.

Due to their heterogeneous origin, the bulk of real-world datasets used for machine learning are very likely to contain missing data and noise. Data mining methods would not produce high-quality results when applied to this noisy data because they would be unable to successfully find patterns. Hence, data processing is crucial to raising the general level of data quality.

We have performed the following preprocessing on the data:

1. **Remove punctuations:** We will be using regular expressions and the ‘re’ library in Python.
2. **Remove the stop words**: Stop words are those words that are frequently used in communication and thereby do not have either a positive or negative impact on our statement such as: “is, this, us, etc.” they do not convey any useful meaning and so they can be directly removed. Hence letters from b to z, will be added to the list of stop words imported directly. We will be using the Natural Language Toolkit (NLTK) library in Python.
3. **Stemming and lemmatization**: Stemming and lemmatization are two techniques used in data preprocessing to reduce words to their base or root form. Stemming is the process of reducing a word to its base form by removing any suffixes. This is often done using a set of predefined rules, such as the Porter Stemmer algorithm. Stemming can result in the same stem for different words, even if they have different meanings. Lemmatization, on the other hand, is the

process of reducing words to their base form, or lemma, based on their morphological analysis. We will be using the Natural Language Toolkit (NLTK) library in Python for this.

4. **Apply counter vectorizer**: To convert a string of words into a matrix of words with column headers represented by words and their values signifying the frequency of occurrence of the word Count Vectorizer is used. We will be using the CountVectorizer from scikit-learn in the python library to apply it in data preprocessing.

5. Work Plan

Project Github Repository:

https://github.com/shivakumar96/Social-Media_Data-Mining.git

Project Name	**Unmasking cyberbullying: Classifying comments and revealing communities**
Project Team	1. Shivakumar Suresh 2. Sanjay Kumar Thovinakere Srinivas 3. Sujay Vishwanath Malghan 4. Rakshitha Kandavara Jayarama 5. Chaitra Sampathram 6. Bindushree Huruhihakkalu Ambikanath 7. Deepika Nandan 8. Riya Jomy Kannampuzha
Team Lead	Shivakumar Suresh
Project Deliverable	28-Apr-2023
Start Date	27-Mar-2023
End Date	27-Apr-2023
Overall Progress	20%
Project Repository (Github)	https://github.com/shivakumar96/Social-Media_Data-Mining.git

Tasks	Responsible	Start	End	Days	Status	Day-1	Day-2	Day-3	Day-4	Day-5	Day-6	Day-7	REMARKS/COMMENTS
SPRINT-1 03/27/2023 - Project proposal													
Set kick-off meeting	Shivakumar	27-Mar-2023	27-Mar-2023	1	Complete								Completed as per scheduled timeline
Create a Proposal Doc	Shivakumar, Sanjay, Sujay, Rakshita, Riya, Bindu, Chaitra, Deepika	28-Mar	2-Apr	6	Complete								Completed as per scheduled timeline
SPRINT-2 04/03/2023 - Data collection and model selection													
Collect Twitter Data	Rakshita, Riya	3-Apr	3-Apr	1	In progress								Twitter data limitations
Data Cleaning	Chaitra, Deepika	4-Apr	4-Apr	1	Not started								
Feature Extraction	Sujay	5-Apr	5-Apr	1	Not started								
Calculate Prior Probablities	Bindu	6-Apr	7-Apr	2	Not started								
Calculate conditional probabilities (Naive Bayes Model)	Sanjay	6-Apr	7-Apr	2	Not started								
Train and Test the Model	Sanjay,Rakshita, Riya, Bindu	8-Apr	9-Apr	2	Not started								
Build a Classifier using Neural Networks Model (Training)	Shivakumar, Sujay	8-Apr	9-Apr	2	Not started								
Validate and Fine-Tune the Model (Testing)	Shivakumar, Deepika, Chaitra	9-Apr	9-Apr	1	Not started								
SPRINT-3 04/10/2023 - Community Detection and Analysis													
Data Visualization Post Classification - 1	Shivakumar, Sujay, Riya, Bindu	10-Apr	12-Apr	3	Not started								
Data Visualization Post Classification - 2	Sanjay, Rakshita, Deepika, Chaitra	10-Apr	12-Apr	3	Not started								
Calculate the Malignancy Levels for Each User	Sanjay	13-Apr	15-Apr	3	Not started								
Create a malignancy-based social network using reciprocal friendships	Shivakumar	13-Apr	15-Apr	3	Not started								
Social Network (community) Visualization	Rakshita, Riya, Bindu, Deepika, Chaitra	16-Apr	16-Apr	1	Not started								
SPRINT-4 04/17/2023 - Final Integration and testing													
Building Hybrid classification Model	Sanjay, Rakshita, Shivakumar	17-Apr	19-Apr	3	Not started								
Testing the data flow	Sujay, Riya, Bindu	20-Apr	21-Apr	3	Not started								
Integration and system testing	Chaitra, Deepika	22-Apr	23-Apr	2	Not started								
SPRINT-4 04/24/2023 - Documentation													
Final analysis and documenting	Sujay, Sanjya, Shivakumar, Bindu, Chaitra, Deepika	24-Apr	26-Apr	3	Not started								
Create PPT	Rakshita, Riya	26-Apr	27-Apr	2	Not started								
Project delivery		28-Apr	28-Apr										

Tasks

SPRINT-1 03/27/2023 - Project proposal			
Set kick-off meeting	Shivakumar	27-Mar-2023	27-Mar-2023
Create a Proposal Doc	Shivakumar, Sanjay, Sujay, Rakshita, Riya, Bindu, Chaitra, Deepika	28-Mar	2-Apr
SPRINT-2 04/03/2023 - Data collection and model selection			
Collect Twitter Data	Rakshita, Riya	3-Apr	3-Apr
Data Cleaning	Chaitra, Deepika	4-Apr	4-Apr
Feature Extraction	Sujay	5-Apr	5-Apr
Calculate Prior Probablities	Bindu	6-Apr	7-Apr
Calculate conditional probabilities (Naive Bayes Model)	Sanjay	6-Apr	7-Apr
Train and Test the Model	Sanjay,Rakshita, Riya, Bindu	8-Apr	9-Apr
Build a Classifier using Neural Networks Model (Training)	Shivakumar, Sujay	8-Apr	9-Apr
Validate and Fine-Tune the Model (Testing)	Shivakumar, Deepika, Chaitra	9-Apr	9-Apr
SPRINT-3 04/10/2023 - Community Detection and Analysis			
Data Visualization Post Classification - 1	Shivakumar, Sujay, Riya, Bindu	10-Apr	12-Apr
Data Visualization Post Classification - 2	Sanjay, Rakshita, Deepika, Chaitra	10-Apr	12-Apr
Calculate the Malignancy Levels for Each User	Sanjay	13-Apr	15-Apr
Create a malignancy-based social network using reciprocal friendships	Shivakumar	13-Apr	15-Apr
Social Network (community) Visualization	Rakshita, Riya, Bindu, Deepika, Chaitra	16-Apr	16-Apr
SPRINT-4 04/17/2023 - Final Integration and testing			
Building Hybrid classification Model	Sanjay, Rakshita, Shivakumar	17-Apr	19-Apr
Testing the data flow	Sujay, Riya, Bindu	20-Apr	21-Apr
Integration and system testing	Chaitra, Deepika	22-Apr	23-Apr
SPRINT-4 04/24/2023 - Documentation			
Final analysis and documenting	Sujay, Sanjya, Shivakumar, Bindu, Chaitra, Deepika	24-Apr	26-Apr
Create PPT	Rakshita, Riya	26-Apr	27-Apr
Project delivery		28-Apr	28-Apr

6. Conclusion

To summarize, the project's goal is to create a classification model that can distinguish various sorts of harmful comments in online interactions and prevent online bullying. The model will be trained on a labeled dataset to identify patterns by leveraging machine learning techniques. The model's implementation can help identify potentially dangerous remarks before they are posted, assist moderators in more efficiently assessing warned comments, and build safer online places by encouraging courteous and constructive dialogues.

In Addition, we make use of network analysis to identify communities behind online bullying. We have developed a comprehensive framework that effectively identifies instances of cyberbullying and uncovers communities with varying degrees of cyberbullying prevalence. Our findings not only facilitate the identification of areas that may require intervention or support but also provide valuable insights for policymakers, educators, and social media platforms to develop targeted and effective strategies to combat cyberbullying. By understanding the behavior patterns and interactions within communities, we can better tailor preventative measures and support systems that address the unique needs and challenges faced by different groups.

Moreover, the proposed framework can be extended and adapted to analyze other forms of online harassment or negative behavior, making it a versatile tool for promoting a safer and more inclusive online environment. As social media platforms continue to evolve and expand, it is crucial for researchers, practitioners, and platform developers to collaborate in the ongoing effort to address cyberbullying and its consequences.

In the future, we recommend further exploration of the relationships between community dynamics and the prevalence of cyberbullying, as well as the integration of additional features, such as user demographics and temporal patterns, to enhance the accuracy and effectiveness of the detection methods. Ultimately, the goal is to foster a positive online ecosystem where users can engage in meaningful interactions without the fear of being subjected to cyberbullying or harassment.

References

- V. Jain, V. Kumar, V. Pal and D. K. Vishwakarma, "Detection of Cyberbullying on Social Media Using Machine learning," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1091-1096, doi: 10.1109/ICCMC51019.2021.9418254.
- W. N. Hamiza Wan Ali, M. Mohd and F. Fauzi, "Cyberbullying Detection: An Overview," 2018 Cyber Resilience Conference (CRC), Putrajaya, Malaysia, 2018, pp. 1-3, doi: 10.1109/CR.2018.8626869.
- N. Tsapatsoulis and V. Anastasopoulou, "Cyberbullies in Twitter: A focused review," 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Larnaca, Cyprus, 2019, pp. 1-6, doi: 10.1109/SMAP.2019.8864918.
- Shetty, J., Chaithali, K.N., Shetty, A.M., Varsha, B., Puthran, V. (2021). Cyber-Bullying Detection: A Comparative Analysis of Twitter Data. In: Chiplunkar, N., Fukao, T. (eds) Advances in Artificial Intelligence and Data Engineering. Advances in Intelligent Systems and Computing, vol 1133. Springer, Singapore. https://doi.org/10.1007/978-981-15-3514-7_62
- Mukhopadhyay, D., Mishra, K., Mishra, K., Tiwari, L. (2021). Cyber Bullying Detection Based on Twitter Dataset. In: Joshi, A., Khosravy, M., Gupta, N. (eds) Machine Learning for Predictive Analysis. Lecture Notes in Networks and Systems, vol 141. Springer, Singapore
- J. O. Atoum, "Cyberbullying Detection Through Sentiment Analysis," 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2020, pp. 292-297, doi: 10.1109/CSCI51800.2020.00056.