IMT575 Data Science III
Professor Caliskan
June 5, 2024

Unimodal Cyberbullying Detection and NLP Analysis

Final Project Report

Shan Ming Gao, CP Chan, Melody Chang, Kelly Liu

1. Motivation and Problem Definition

Cyberbullying as defined by the American Psychological Association is the "willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices." (Abramson, 2022)[1]. Being a victim of cyberbullying is associated with depression, anxiety, suicidal thoughts and attempts, and more. In 2023, a study of 26.5% of 5,000 nationally representative, middle and high schoolers reported being bullied in the last 30 days; this statistic has also risen, from 23.2% in 2021 and 17% in 2019 (Hinduja and Patchin, 2024).

Thus, cyberbullying detection, understanding, protection, and prevention *at scale* are highly valuable for public mental health. Our data science project contributes to the foundational detection and understanding components.

Problem Statement: How may we create an unimodal cyberbullying detection model that incorporates text analysis in predicting the presence of cyberbullying and identifying the social context associations of the messages?

Project Questions

1. Understand cyber-bullying dataset context by using natural language processing techniques to do text preprocessing and sentiment analysis[2]
2. Cyberbullying detection model (unimodal)[3] to the prediction of if the context is bully or not, and which category the content falls into
3. Test the prediction model's accuracy, precision, recall, and F1 score using the confusion matrix

2. Related work (Directly related)

Cyberbullying detection with machine learning has been practiced before, and deep neural networks in particular are more effective than conventional techniques (Raj et al., 2022)[4]. Previous work on cyberbullying detection has also considered other factors, such as cyberbullying role detection (ie. defenders, bystanders, and instigators), non-textual cyberbullying (eg. with OCR classification) (Logasree & Harshini, 2023)[5], and personality associations (Balakrishnan et al, 2020)[6].

---

[1] Hinduja, S. & Patchin, J. W. (2024). Cyberbullying fact sheet: Identification, Prevention, and Response. Cyberbullying Research Center. Retrieved May 17, 2024, from https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2024.pdf

[2] Twitter sentiment analysis: Nitin G. (2019, April 3). Twitter Sentiment Analysis—Word2vec, doc2vec. Kaggle. https://kaggle.com/code/nitin194/twitter-sentiment-analysis-word2vec-doc2vec

[3] Civis Analytics. (2018, March 8). An Intro to Natural Language Processing in Python: Framing Text Classification in Familiar Terms. The Civis Journal. https://medium.com/civis-analytics/an-intro-to-natural-language-processing-in-python-framing-text-classification-in-familiar-terms-33778d1aa3ca

[4] Raj, M., Singh, S., Solanki, K., & Selvanambi, R. (2022). An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques. Sn Computer Science, 3(5), 401. https://doi.org/10.1007/s42979-022-01308-5

[5] Logasree, S., & Harshini, M. (2023). Cyberbullying Detection using machine learning. International Research Journal of Education and Technology. https://www.irjweb.com/Cyberbullying%20Detection%20using%20machine%20learning.pdf

[6] Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. Computers & Security, 90, 101710. https://doi.org/10.1016/j.cose.2019.101710

3. Methodology (Data, Transformation, and Evaluation) & Results

Our work is made possible by the efforts of Ahmadinejad et al. (2023)[7] who generated a dataset of 99,991 Tweets (Twitter/X), with one label for non-cyberbullying and three for cyberbullying of racial, religious, or gender-and-sexuality nature. Significantly, their labels were verified by randomly sampling 1000 tweets and giving them to three "social media specialists with experience in detecting cyberbullying" - ie. domain experts - who affirmed over 90% classification accuracy.

Among all metrics, *Recall* will be the most important metric for us, the less false negative the better. We do not want our model to see a tweet that is supposed to be cyberbullying content and think that it is not. The feature input will be the word embedding which is converted from the raw tweet.

For the text preprocessing, we use Natural Language Toolkit NLTK -Tokenization and Stopwords to remove common words lacking significant meaning. Then we load the pre-trained word2vec model from Google News with gensim. downloader. After having the pre-trained word2vec model, we convert the raw tweet to word embeddings, which is the numeric vector input. The distance and direction between vectors indicate the similarity and relationships between words. This will be the feature we used for all the detection models.

A. Text and Sentiment Analysis

In the text-cleaning phase, we will create four lists based on the categories on the label. Following categorization, we will conduct frequency analysis to extract insights from the most commonly occurring words and contexts within each category. Additionally, we will use word clouds for visual representation to clearly show the "keywords" in each category.

Figure 1. Word Clouds by Category

Note: Due to the nature of our topic, many of the words below are highly offensive.

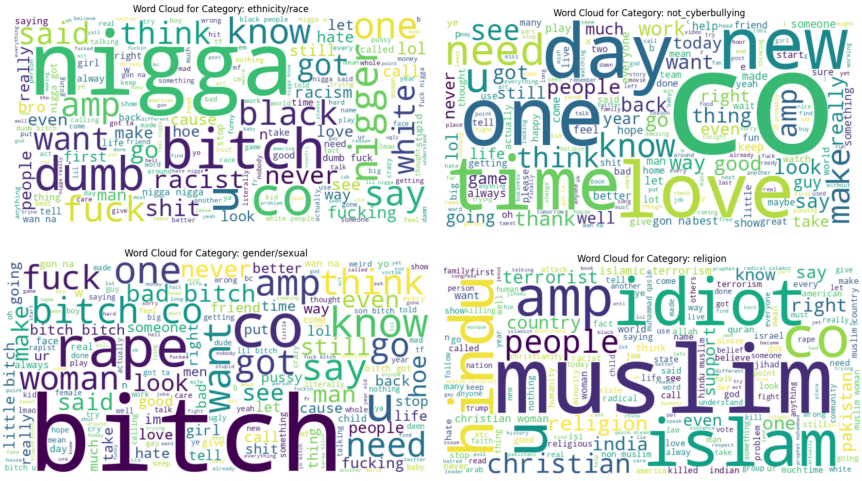

Table 1. Sentiment Analysis Results

[7] Ahmadinejad, Mohamadreza, Shahriar, N., & Fan, L. (2023). Self-Training for Cyberbully Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset.
https://www.semanticscholar.org/paper/Self-Training-for-Cyberbully-Detection%3A-Achieving-a-ahmadinejad-Shahriar/2ca8c44f6f3e0a9f740744f99dff557879eb6279

Category: ethnicity/race Polarity score: -0.08 Sentiment: Negative	Category: gender/sexual Polarity score: -0.04 Sentiment: Negative
Category: Religion Polarity score: -0.03 Sentiment: Negative	Category: not cyber bullying Polarity score: 0.15 Sentiment: Positive

B. Cyberbullying Detection Model

1. MLPClassifier

We chose Multi-Layer Perceptrons (MLP) as one of the models for cyberbullying detection. For the hidden layers, we chose 4 hidden layers {200,100,50,25}, with the layers from bigger to smaller size. The reason that we set the hidden layers this way, is that with a bigger size on the first layer, and a gradually smaller size, it can help the network learn a hierarchy of features, where more features are captured in the wider layers and more detailed features in the narrower layers.

From the results, it has a high Recall value, indicating that MLPClassifier is doing a good job of cyberbullying detection. Based on the Recall, this MLPClassifier model performed well on “Not cyberbullying” and “ethnicity/race” content. 0.98 for “Not cyberbullying” in Recall, and 0.98 for “ethnicity race”. [Appendix Table 1]

2. TensorFlow

For the TensorFlow, we start with the label encoder which turns categorical labels into numerical labels. After splitting training and testing sets, we can build the sequential model by adding dense and dropout layers to prevent overfitting. We can use test loss to see how well a machine learning model performs on data that it has not seen before, and identify whether it is overfitting or underfitting.

The model has a low test loss, indicating that it is not overfitting or underfitting. It also has a high Recall value just like the MLPClassifier, indicating that it is doing a good job on cyberbullying detection. We then can look into the performance of each cyberbullying and non-cyberbullying category. Based on the Recall, this Tensorflow model performed well on “Not cyberbullying” and “religion” content. 0.99 for “Not cyberbullying” in Recall, and 0.97 for “religion”. [Appendix Table 2]

3. Logistic Regression

The model is set with the solver as 'Liblinear' to handle highly sparse data and includes regularization to prevent overfitting on the training data. It shows better precision overall as well as classifying and identifying 'Non-cyberbullying' contents. However, it struggles the most when it comes to classifying positive instances within the 'ethnicity/race' category. [Appendix Table 3]

4. SVM

SVM operates effectively in N-dimensional space and can handle non-linear relationships. The kernel setting for the SVM classification model is RBF, since its superior performance in text or image categorization tasks. Results indicate the model can almost perfectly categorize religion-related bullying content and reliably detect non-cyberbullying content. However, instances of bullying based on ethnicity/race are more likely to be overlooked by the model. [Appendix Table 4]

4. Key Findings and Conclusion

Sentimental Analysis: Based on polarity scores, all cyberbullying content exhibits a negative sentiment, with the ethnicity/race group showing the most negativity, while non-cyberbullying content receives a positive sentiment.

Text Analysis: Black people and women are major targets of cyberbullying. Terms like the "n-word" and the "b-word" frequently surface in the ethnicity/race category. These two groups have been historically marginalized, and unfortunately, this reality persists today. Insults aimed at women appear in the gender/sexual category as well. We see the "b-word" once again, alongside "women," "rape," and the "f-word", highlighting women continue to be viewed as vulnerable targets of disrespect and crime. In the religion category, 'Muslim' is at the top of the cyberbullying list, illustrating the lasting impact of 911. In non-cyberbullying content, the most frequent words are 'love' and 'time', with many of the words reflecting neutral or positive sentiments.

Prediction Models: MLP Classifier, Tensorflow, and SVM all achieved high scores of 0.98 across accuracy, precision, recall, and F1 Score. On the other hand, Logistic Regression achieved slightly lower scores of 0.94 across all metrics, this could be due to it works on linear relationships only.

Two other key findings observed, first, all the models have stronger ability to capture non-cyberbullying content as indicated by higher recall values (Table 2). This could be because of imbalance dataset, there are much more data on non-cyberbullying content than on other types. Secondly, most models struggle to identify all instances of bullying content related to ethnicity or race, as it has the lowest recall values in general. This could be because many keywords overlap with gender/sexual, making it harder for models to detect.

Table 2. Classification Report of Recall by Model

	MLPClassifier	TensorFlow	Logistic Regression	SVM
Overall	0.98	0.98	0.94	0.98
Ethnicity/Race	0.98	0.96	0.88	0.95
Gender/Sexual	0.95	0.98	0.90	0.96
Religion	0.97	0.97	0.94	0.97
Not_Cyberbullying	0.98	0.99	0.97	0.99

5. Ethical Limitations and Future Work

There are several limitations to the model and analysis of our work, which also suggest potential future directions. One major limitation is *unimodality*; social media content often extends beyond text, encompassing images, videos, and interactions that are not captured in isolated tweets or words. Additionally, the dataset's nearly 100k tweets are better balanced as two classes (cyberbullying or not; vs 9916 vs 10082 observations, respectively) than when analyzed as four classes due to the dataset’s

structure. However, data imbalances are also reflected in real-world scenarios, where only about 10% of tweets encountered by Ahmadinejad et al. (2023) were classified as cyberbullying. Moreover, the dataset's limitation to four classes oversimplifies the complexity of real-world behavior where topics are not mutually exclusive and often overlap. This especially implicates the concept of *intersectionality* in multi-topic cyberbullying: for instance, a Tweet targeting Black women would need to be classed race *or* gender (not both), reducing the accurate representation of *misogynoir*[8] in the real world. Furthermore, the three cyberbullying classes do not cover all legally protected class[9] categories, such as disability, age, and political orientation, with *user age* being particularly crucial in the context of cyberbullying. Future work could expand on existing research by incorporating classifications based on cyberbullying roles, psychometric and personality data, and user behavior to create a more comprehensive and nuanced understanding of cyberbullying on social media.

[8] Bailey, M., & Trudy. (2018). On misogynoir: citation, erasure, and plagiarism. *Feminist Media Studies,* 18(4), 762–768. https://doi.org/10.1080/14680777.2018.1447395

[9] (Not accounting for state-differences) California, for instance: https://www.senate.ca.gov/protected-classes

Appendix

Table 1. MLPClassifier Multiclass Classification

Result

- Accuracy: 0.9756475647564756

	Precision	Recall	F1	Dataset
Overall	0.98	0.98	0.98	19998
ethnicity/race	0.95	0.98	0.97	3425
gender/sexual	0.99	0.95	0.97	3336
religion	0.97	0.97	0.97	3155
Not cyberbullying	0.98	0.98	0.98	10082

Table 2. Tensorflow Multiclass Classification

Result:

- Test Loss: 0.09214311093091965
- Accuracy: 0.9785478547854786

	Precision	Recall	F1	Dataset
Overall	0.98	0.98	0.98	19998
ethnicity/race	0.97	0.96	0.97	3425
gender/sexual	0.97	0.98	0.97	3336
religion	0.99	0.97	0.98	3155
Not cyberbullying	0.98	0.99	0.98	10082

Table 3: Logistic Regression Classification Report:

Accuracy: 0.9371437143714372

	Precision	Recall	F1-Score	Dataset
Overall	0.94	0.94	0.94	19998
Ethnicity/Race	0.95	0.88	0.91	3425
Gender/Sexual	0.94	0.90	0.92	3336
Not_Cyberbullying	0.93	0.97	0.95	10082
Religion	0.96	0.94	0.95	3155

Table 4: SVM Classification Report

Accuracy: 0.9758475847584759

	Precision	Recall	F1-Score	Dataset
Overall	0.98	0.98	0.98	19998
Ethnicity/Race	0.97	0.95	0.96	3425
Gender/Sexual	0.97	0.96	0.97	3336
Not_Cyberbullying	0.97	0.99	0.98	10082
Religion	0.99	0.97	0.98	3155