Detecting Cyberbullying through Tweets and Network Graphs

Yamini Ranganathan, Rachel Porter, Yashpreet Malhotra
DSCI6004 Final Project Report
University of New Haven

Abstract: Explore the possibility of an increase in cyberbullying incidents due to the pandemic and high social media usage by creating a model that can automatically successfully detect instances of bullying tweets. To successful report/detect instances of cyberbullying in social media three different models, a Naïve Bayes Classifier, Bi-directional LSTM, and Bert Transformer were evaluated for Twitter Sentiment Analysis. Network graphs were used to identify the patterns of relationship between different cyberbullying categories. Among the models evaluated Bert Transformer with additional dense layer showed the maximum overall accuracy of 95% with F1 scores over 95%. Both parallel coordinates plot graphs and network graphs indicated that cyberbullying can be considered as a negative post in multiple categories.

Key words: Cyberbullying, Tweets, Sentiment Analysis, Naïve Bayes, Bi-Directional LSTM, Bert Transformer, Network Graphs.

GitHub link: https://github.com/yashmalhotra124/DSCI-6004---NLP/

Introduction:

Social media is a huge force in today's world, and seemingly only continues to grow. Platforms like Facebook, Twitter and Instagram have become part of the majority of people's daily routines. This is even more true for younger generations who have been raised with social media. While social media is great for a lot of reasons, it also makes the bad aspects of socializing more accessible too. Cyberbullying has become a large issue with the rise of social media, leading to increasing mental health problems in today's youth. Twitter and Facebook have some of the highest rates of bullying across all platforms.

This became even more true during the 2020 COVID-19 pandemic, where a 20% increase in social media usage was observed. In April 2021, an AI company dedicated to monitoring hate speech online noted a 70% increase in the amount of hate speech among teens and children in online chats (Security.org Team, 2022).

Covid pandemic led to an uprise of digital presence, which many hypothesize has led to a rise of cybersecurity attacks, including cyberbullying. Karmakar and Das (2021) analyzed the tweets from January 1st, 2020 - June 7th, 2020 and their Bayesian method showed an upward trend on cyberbullying-related tweets since mid-March 2020 indicating a correlation of the crisis with the discussion of such incidents by individuals, Bartlet et al., (2021) reported an significant increases in BIMOB, cyberbullying attitudes, and cyberbullying perpetration during the pandemic resulting from their study of the cyberbullying frequency during COVID-19 pandemic by comparing US adult participant data from six months before the start of the pandemic to the middle of the pandemic.

How does one combat this issue? Detection of cyberbullying and the subsequent preventative measures taken after identifying instances are the main courses of action in ending and preventing cyberbullying. In this paper, the aim is to provide an effective method of detecting cyberbullying in tweets from Twitter. By doing so, it will be possible for individuals, parents, caregivers, and medical professionals to analyze social media posts for linguistic clues that signal deteriorating mental health far beyond traditional approaches, and stop the cyberbullying at the source.

User interaction data from Twitter—including likes, follows, and retweets—provides valuable information what kinds of discussions are taking place online, and who is participating in them (Korobova & Zacharski, 2021). This data can be represented using graphs/networks, that users can interact with and analyze the trends and relationships.

Review of Literature: The state of the art and relevant works on the use of machine learning and deep learning models used to deduct cyber bullying are presented. NLP techniques & supervised machine learning was used to detect cyberbullying by not considering sarcastic content as cyberbullying but by identifying the themes or categories associated with cyberbullying (Pereraa & Fernandob, 2020). Automatic detection of cyberbullying in social media text by identifying different types of bullying (Van Hee et al., 2018). A Hybrid Model for Detecting Cyberbullying in the Spanish Language by creating different models that combine the Lexicons and ML approach (Lepe-Faúndez et al., 2021). A Multichannel Deep Learning Framework for Cyberbullying Detection on social media uses the bidirectional gated recurrent unit (BiGRU), transformer block, and convolutional neural network (CNN), to classify Twitter comments into two categories: aggressive and not aggressive (Munif Alotaibi et al., 2021). SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection focuses on the ability to discern what specific quality of the victim the cyberbully is attacking (Wang et al., 2020)

Materials and Methods:

2021 Fine-Grained Balanced Cyberbullying Dataset: The dataset used was balanced and separated into tweets that fall into categories of cyberbullying targeting a victim's age, ethnicity, gender, religion, or other quality.

Link to dataset: https://ieee-dataport.org/open-access/fine-grained-balanced-cyberbullying-dataset

Methods:

First, the dataset was loaded and checked for duplicates, and class balance. There were 36 duplicates that were dropped, and each class had roughly 7900 samples. The tweets were then "cleaned" by removing emojis, punctuation, special characters (&, $, etc.) and multiple sequential spaces by specially defined functions. After tweets were cleaned, they were again

checked for duplicates, resulting in another 3,058 tweets being removed from the dataset. After doing so the classes were still balanced except for the “other_cyberbullying” class, so it was decided to remove this class and related tweets.

Another column in the dataframe of the dataset was created with the length of the cleaned tweet. Tweets shorter than 4 words and more than 100 words were removed from the dataset, leaving 37,113 tweets. Once the tweets were cleaned and processed, the “sentiment” column in the dataframe was created which encoded ‘religion’ to 0 and ‘age’ to 1, etc.

The dataset was split into training and test sets and then training was split again into training and validation sets. When looking at the resulting classes in the training set, they had become slightly unbalanced, so the training set was oversampled to ensure that all classes had the same number of tweets as the most populated one.

The first model created was a Naive Bayes baseline model. A bag of words was created using CountVectorizer and a term-frequency times inverse document frequency (TF-IDF) transformation was performed on both the training and test sets--this is a common term weighting scheme in information retrieval as it associates weights to the different words based on their frequency such that rarer words will be given more importance.

Once this was done, a Naive Bayes model was instantiated and fitted to the training data and tested on the test set. Results will be discussed in the Results section of this paper.

In addition to the Naive Bayes model, a custom Bidirectional LSTM model was also created. Here the preprocessing involved tokenizing the sentences with a custom defined function. The sentences are converted to lists of numbers and are padded to the max sentence length.

The top 20 most common words were extracted from the vocabulary dictionary that was created with the tokenizer function. A word embedding matrix was then created using the original text tweets and the pre-trained model Word2Vec. The data was again split into training and test, and the training set was oversampled to ensure all classes had the same number of samples as the most populated class.

The PyTorch LSTM model was then created, with the following specs:
Number of classes: 5
Hidden dimensions: 100
LSTM layers: 1
Learning rate: 3e-4
Bidirectional: True
Epochs: 5

The batch size was set to 32 and the model was trained for 5 epochs. Results will be discussed in the Results section of this paper.

Another model approach considered was BERT. A pre-trained BERT model was loaded fom the Hugging Face library and fine-tuned for the classification task. Again the data was split into training, validation and test, and RandomOverSampler() was again applied to ensure that the classes were balanced.

A custom tokenizer was created for loading the tweets. In order to specify the length of the longest tokenized sentence, the train tweets were tokenized using the "encode" method of the original BERT tokenizer and the length of the longest sentence was checked.

The pre-processing for the BERT model was similar to the other models. The arrays and target columns were converted to tensors. Then a dataloader was created for fine-tuning the BERT model.

A custom BERT classifier class was created, including the original BERT model made of transformer layers, and additional Dense layers to perform the desired classification task. A Linear Warmup rate as used--which is a learning rate schedule where the learning rate is linearly decreased from a low rate to a constant rate thereafter. This reduces volatility in the early stages of training. We defined a function that creates a schedule with a learning rate that decreases from the initial set lr in the optimizer to 0, after a warmup period during which it increases linearly from 0 to the initial lr set in the optimizer. The optimizer used was AdamW which yields better training loss and better generalization than the models trained with Adam. Cross entropy was the loss function.

The model was trained for 2 epochs over 1,775 batches. A function called bert_predict was created to return a list of the predictions from the test dataset.

In addition to the above models, a cyberbullying tweets network graph and analysis was performed to analyze the top frequency words and most used phrases in the tweets. This was visualized using a Parallel Coordinates Plot (PCP) which is a visualization technique used to analyze multivariate numerical data and for comparing many quantitative variables together by looking for patterns and relationships between them.

The tweets were split into separate datasets based on the specific criterion of cyberbullying, i.e. gender, religion, ethnicity, age, and not cyberbullying as seen in the original dataset. When preprocessing the data for the PCP, the datasets were created and an inner merge (intersection)

was performed on the gender and religion datasets. Then a left join was done on the remaining datasets keeping the Text column as the common column. Nan values were replaced with 0s.

In addition to the analysis of the tweets, the network graph as created which is used to display relationships between elements (nodes) using simple links. Network graphs allow us to visualize clusters and relationships between the nodes quickly. To create the network of words, generate word pairs as bigrams. Till now, unigrams were used for plots and analysis. Generate bigrams of those words which carry the main weight or sentiment of the whole text body. In simpler words, they are those word phrases which are giving the whole text an bullying sentiment.

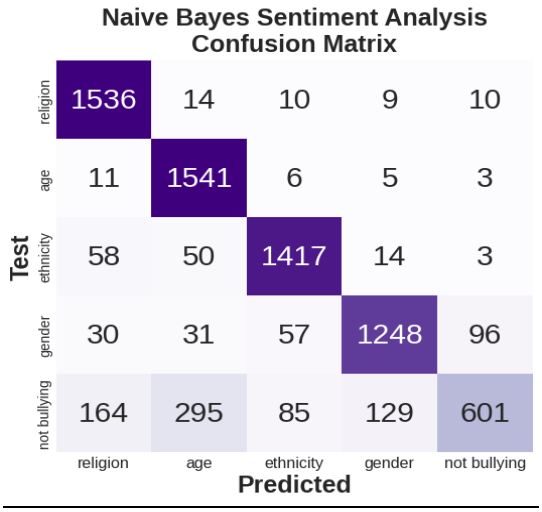
Bigrams might contain more meaning and weight as compared to unigrams and are used for network analysis. Also as observed in the Parallel Coordinate plot above that many unigrams are occurring for multiple types of cyberbullying category. For example, the word women is taken into account for multiple cyberbullying type categories such as gender and religion. To make it quick and easy to observe this type of relation and how many words might be interconnected within or across multiple categories, so as to create network plots as below.

By using a DiGraph class from the network, it is the base class for directed graphs and can store nodes and edges with many optional data. The degree of each node is the number of edges adjacent to the node. The nodes in the plots will be colored according to the size of the node.

Results:

Naive Bayes Classification: Overall accuracy for Naive Bayes was 85%. The Precision, Recall and F1 score for all cyberbullying classes have very high F1 scores, whereas for the class "non-cyberbullying" the score is much lower and the confusion matrix is presented in Figure 1.

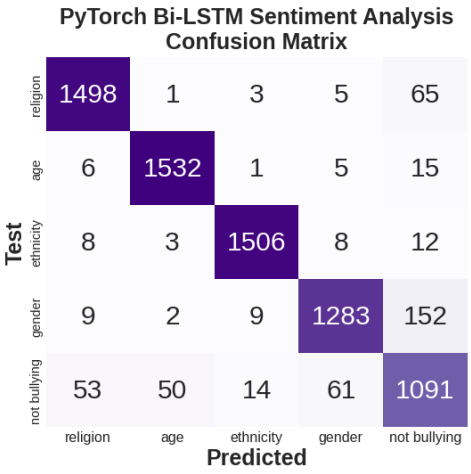
Figure 1: Naïve Bayes Confusion Matrix



Custom Bidirectional LSTM Model: The performance scores had an overall accuracy of 93%. The F1 scores are over 90% for all the cyberbullying classes and the confusion matrix is presented in Figure 2.

Figure 2: Custom Bidirectional LSTM Model Confusion Matrix and Accuracy Score

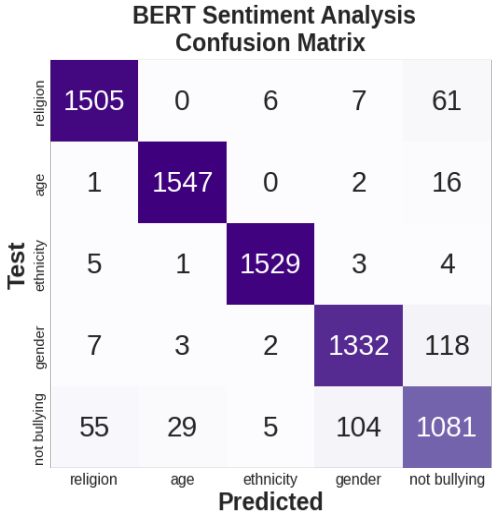
```
Epoch 1:Validation accuracy increased (0.000000 --> 92.153533).  Saving model ...
        Train_loss : 0.6667 Val_loss : 0.2316
        Train_acc : 74.517% Val_acc : 92.154%
Epoch 2:Validation accuracy increased (92.153533 --> 93.240489).  Saving model ...
        Train_loss : 0.1797 Val_loss : 0.1998
        Train_acc : 94.134% Val_acc : 93.240%
Epoch 3:Validation accuracy increased (93.240489 --> 93.410326).  Saving model ...
        Train_loss : 0.1153 Val_loss : 0.1988
        Train_acc : 96.216% Val_acc : 93.410%
Epoch 4:Validation accuracy did not increase
        Train_loss : 0.0750 Val_loss : 0.2256
        Train_acc : 97.601% Val_acc : 92.595%
Epoch 5:Validation accuracy did not increase
        Train_loss : 0.0517 Val_loss : 0.3084
        Train_acc : 98.478% Val_acc : 89.538%
```



BERT Classification: The performance scores of BERT Classifier are quite high and higher than those achieved using the LSTM model with an overall accuracy of 94% and F1 scores over 95% and the confusion matrix is presented in Figure 3.

Figure 3: BERT Model Confusion Matrix and Accuracy Score

```
----------
Epoch : 2
----------
---------------------------------------
BATCH NO. |  TRAIN LOSS  | ELAPSED (s)
---------------------------------------
   100    |   0.149053   |   65.16
   200    |   0.140416   |   64.53
   300    |   0.151965   |   64.49
   400    |   0.178391   |   64.46
   500    |   0.138347   |   64.52
   600    |   0.143743   |   64.47
   700    |   0.110452   |   64.49
   800    |   0.111650   |   64.51
   900    |   0.111064   |   64.52
  1000    |   0.123488   |   64.51
  1100    |   0.150879   |   64.47
  1200    |   0.132064   |   64.50
  1300    |   0.132465   |   64.51
  1400    |   0.111030   |   64.52
  1500    |   0.124794   |   64.55
  1600    |   0.138664   |   64.57
  1700    |   0.128602   |   64.51
  1775    |   0.106077   |   48.33
-------------------------------------------------------------------
AVG TRAIN LOSS |  VAL LOSS  | VAL ACCURACY (%) | ELAPSED (s)
-------------------------------------------------------------------
   0.132777    | 0.198950 |       94.66       |  1185.16
-------------------------------------------------------------------
```



Network Analysis & Graphs: Bar graphs (Figures 4-8) were created for the multiple categories of cyberbullying to analyze word occurrences.

Figure 4: Bar graph for the cyberbullying category, Gender.

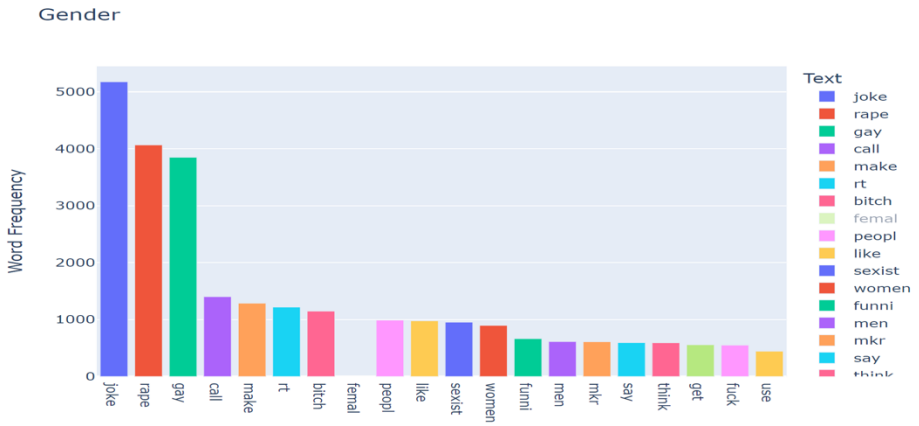


Figure 5: Bar graph for the cyberbullying category, Religion

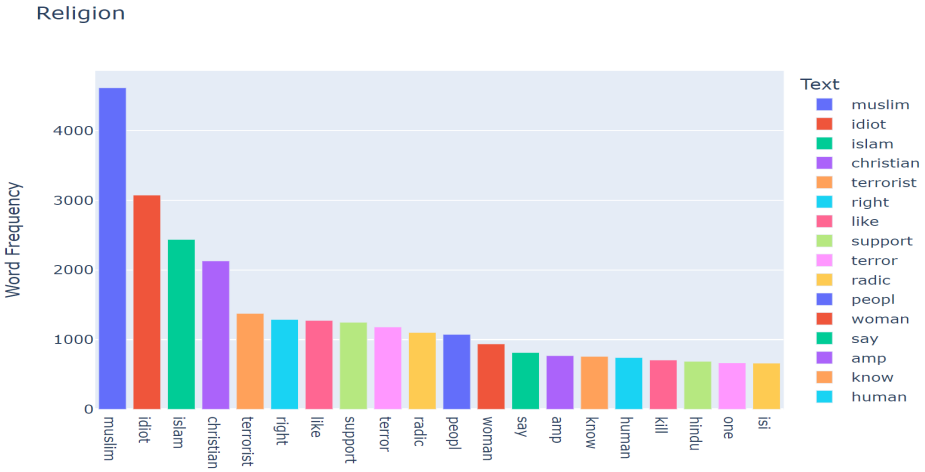


Figure 6: Bar graph for the cyberbullying category, Ethnicity

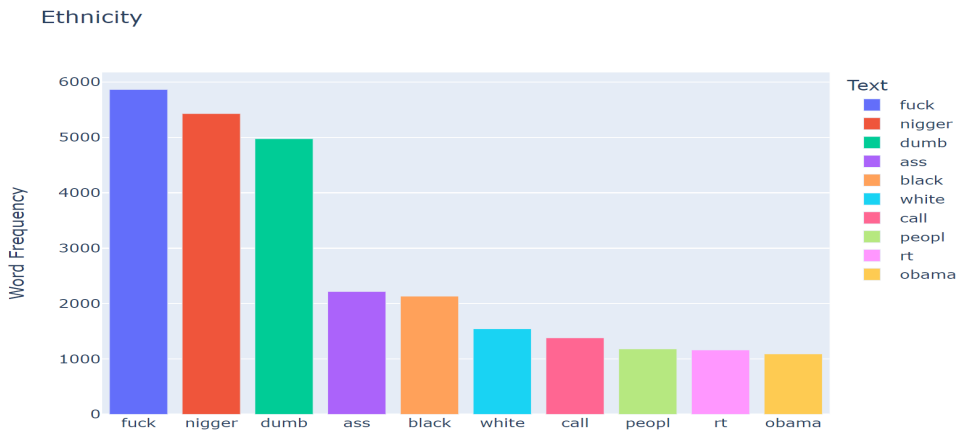


Figure 7: Bar graph for the cyberbullying category, Age

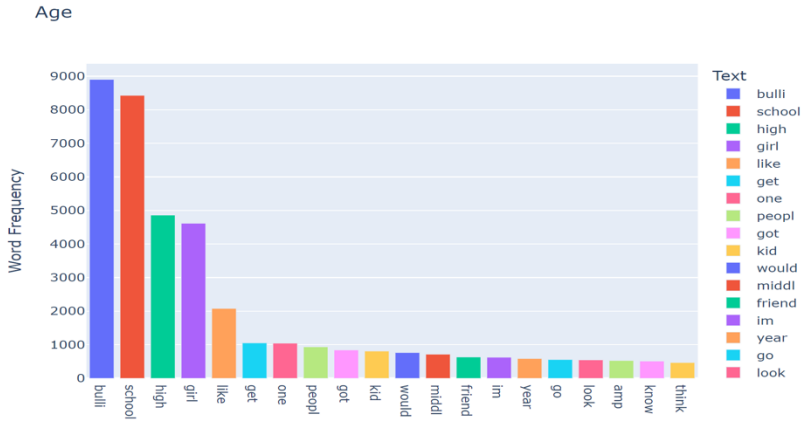
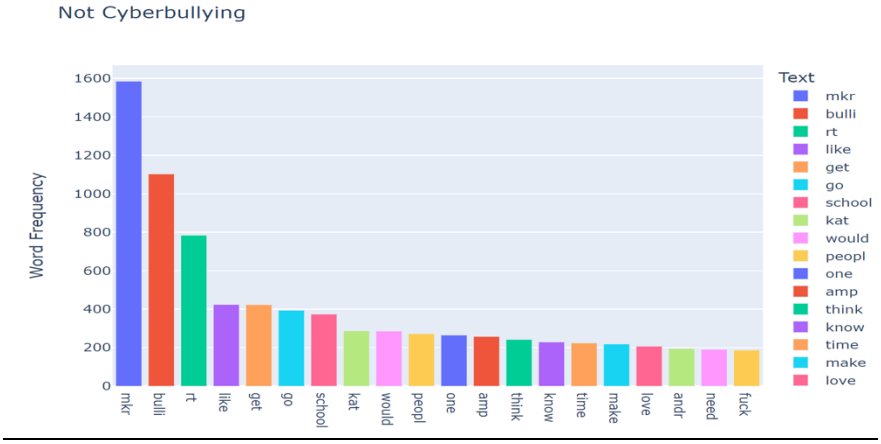
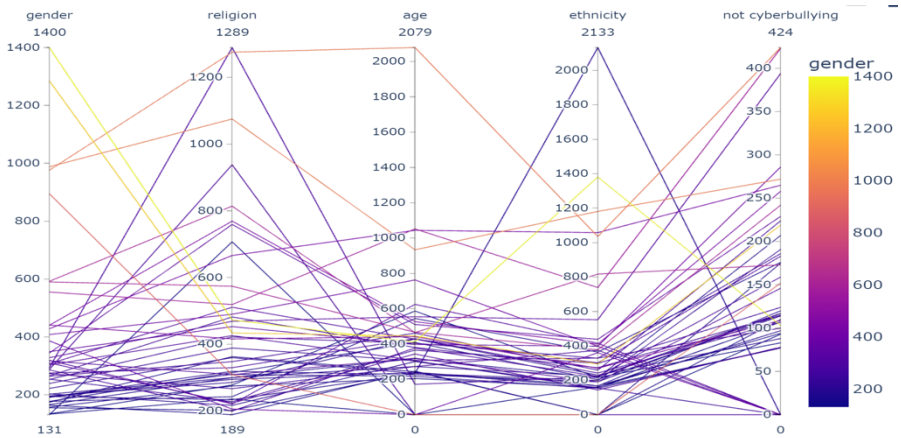


Figure 8: Bar graph for the cyberbullying category, Not-Cyberbullying



The Parallel Coordinates Plot for all the different categories of cyberbullying is presented in Figure 9.

Figure 9: Parallel Coordinates Plot for all cyberbullying categories



Network Graphs: The DiGraph class from the network graphs specify the degree of each node in each category of cyberbullying is depicted in Figures 10-14. For interactive graphs, see our Jupyter notebook.

Figure 10 & 11: Gender network graph (left), Age network graph (right)

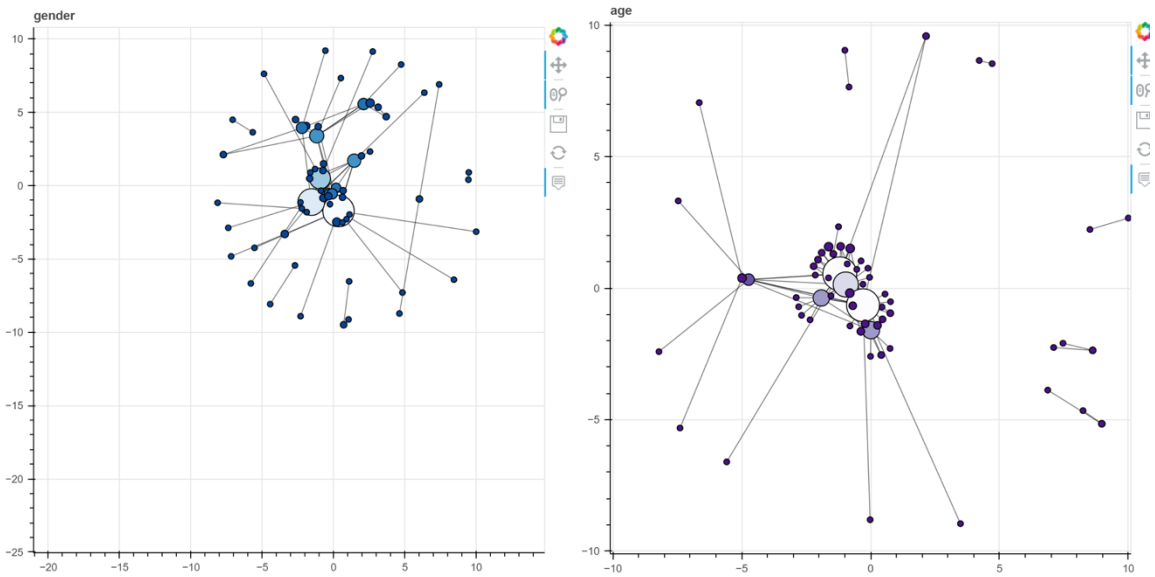


Figure 12 & 13: Religion network graph (left), & Not-cyberbullying network graph (right).

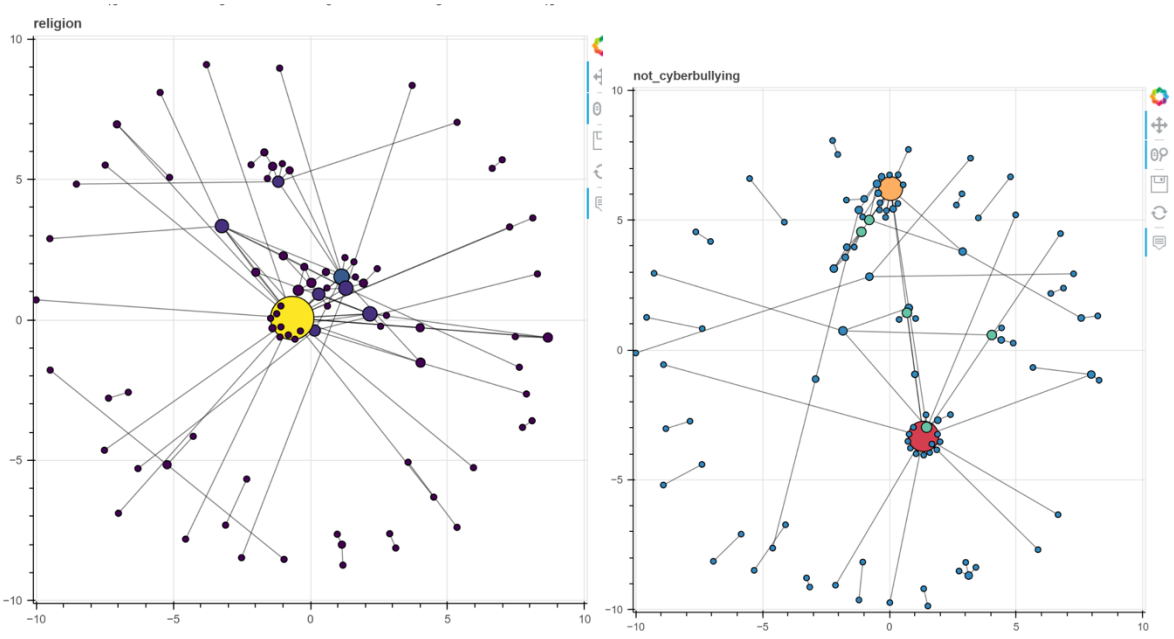
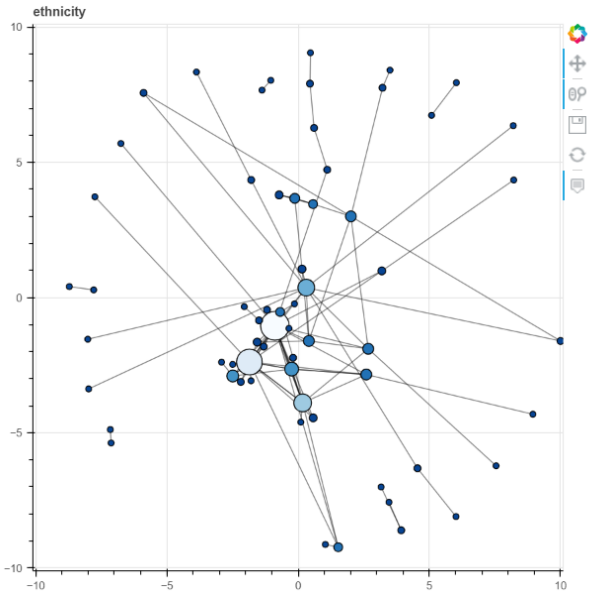


Figure 14: Ethnicity network graph



Discussion:

The Naive Bayes classifier achieved an overall accuracy score of 85%. The precision, recall, and F1 score for all cyberbullying classes were quite high, whereas for the class "non-cyberbullying" the score is much lower. 85% accuracy was not bad, but there was room for improvement. The next model trained was a Custom Bidirectional LSTM. This produced an overall accuracy score of 93%. Additionally, the F1 scores were all over 90% for all of the cyberbullying classes. This is better and quite accurate.

The next model trained was a BERT model that was fine-tuned for the classifier task. The performance scores of this model were quite high and higher than both the Naive Bayes and LSTM model with an overall accuracy of 94% and F1 scores over 95%.

In the analysis of the network and datasets created from the original that were split into categorization of type of cyberbullying, an analysis was conducted on the most frequent words used in the tweets identified as cyberbullying. From the gender category the top 3 words were "joke", "rape", and "gay". From the religion category, the top 3 words were "muslim", "idiot", and "isalm". From the ethnicity category, the top 3 words were "fuck", "ni**er", and "dumb". From the age category, the top 3 words were "bulli", "school", and "high".

From the parallel coordinates plot graph in the Results section, you can see that a lot of cyberbullying posts might not be directed towards just one criterion, but can be thought of or perceived as a negative post in multiple categories.

Even though more students reported that they had experienced recent cyberbullying in 2021 (22.6%) compared to previous years (17.2% in 2019 and 16.7% in 2016), without context (such as knowing the relationship between the aggressor and target and whether the actions were intentionally hurtful and repeated over time) it is difficult to definitively determine if something posted online qualifies as bullying (Patchin, 2021). This is where the network graphs come in handy to identify the context and relationship types.

In the network graphs, you can see some of the same patterns. Looking at the age graph for example, the largest nodes are school, bulli, and girl, and they share a lot of the same connections. In the religion network graph, muslim is the largest node, by 27 degrees--which shows how targeted the nature of these bullying tweets are, and that they seem to be specific to the muslim group, or at least highly concentrated towards them.

Overall, the best model that was trained was the BERT model. It had both high accuracy and high F1 scores. This is most likely due to the underlying architecture of the BERT model. It makes use of transformers, an attention mechanism that learns contextual relations between words or sub-words in a text and learns how important all of the words are in the sentence by looking at their positioning. Models with high accuracy and high F1 score are beneficial and can make online spaces safer not just for kids, but for everyone participating in the use of social media.

References:

1. Alotaibi, M.; Alotaibi, B.; Razaque, A. (2021). A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media. Electronic, 10, 2664.

2. Barlett, C.P., Simmers, M.M., Roth, B., Gentile, D. (2021). Comparing cyberbullying prevalence and process before and during the COVID-19 pandemic. J Soc Psychol. Jul 4;161(4):408-418.

3. Karmakar, Sayar and Das, Sanchari, Understanding the Rise of Twitter-Based Cyberbullying Due to COVID-19 through Comprehensive Statistical Evaluation (January 4, 2021). In Proceedings of the 54th Hawaii International Conference on System Sciences. 2021, Maui, Hawaii (Virtual). https://ssrn.com/abstract=3768839 or http://dx.doi.org/10.2139/ssrn.3768839

4. Korobova, K., and Zacharski, M. (2021). A Network Analysis of Twitter Discourse About The Death of George Floyd. Carleton College. https://www.causeweb.org/usproc/sites/default/files/usresp/2021-1/A%20Network%20Analysis%20of%20Twitter%20Discourse%20About%20The%20Death%20Of%20George%20Floyd.pdf

5. Lepe-Faúndez, M.; Segura-Navarrete, A.; Vidal-Castro, C.; Martínez-Araneda, C.; Rubio-Manzano, C. (2021). Detecting Aggressiveness in Tweets: A Hybrid Model for Detecting Cyberbullying in the Spanish Language. Appl. Sci. 11, 10706.

6. Patchin, J. W. (2021). Bullying During the COVID-19 Pandemic. https://cyberbullying.org/bullying-during-the-covid-19-pandemic

7. Pereraa & Fernandob, 2020, Accurate Cyberbullying Detection and Prevention on Social Media, CENTERIS.

8. Security.org Team. (2022). Cyberbullying: Twenty Crucial Statistics for 2022. https://www.security.org/resources/cyberbullying-facts-statistics/

9. Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PloS one, 13(10).

10. Wang, J., Fu, K., Lu, C.T. (2020). SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection. Proceedings of the 2020 IEEE International Conference on Big Data (IEEE BigData 2020), 1699-1708, December 10-13.