# A novel approach to classifying Cyberbullying Tweets

Srinivas Akhil Mallela
Hemanth Chenna

University of Colorado Boulder

***Abstract*–During the past few years of the pandemic, cyberbullying has become an even more serious threat. Our work aims to devise a set of models to outperform previous related work on this problem. Previous literature has not yet explored XLNet and ColBERT models for this use case and we establish some experiments to explore the viability of these models, which were previously used for tasks such as humor detection, to identify cyberbullying tweets from Twitter. Our models are compared against the baseline model used in prior literature of the SOSNet paper, which was evaluated on the same dataset. From our results, we observed that our models are able to classify the tweets well and show promise for further improvement but still fall slightly short of the baseline model overall, although they outperform the baseline in some classes.**

## 1 Introduction

The onset of the pandemic led to more people spending time in increasing degrees of isolation, which in turn forced people to turn to the internet to achieve the basic day-to-day social interaction. This allowed them to become easy targets for malicious actors online and subsequently experiencing the harmful effects of online cyberbullying. It has now become a cause for concern among social media companies and the community to curb this unfortunate side-effect of the pandemic. We aim to contribute to solving this problem with a novel approach that we believe will be more effective than the prior work.

All of the prior work we have come across, have experimented with classic machine learning, neural network and natural language processing models such as Naive Bayesian models, BiLSTM, Random Forests, and using Word2Vec and BERT embeddings in neural networks. We believe there are some modern approaches which perform better than these classical methods. We have listed some of these below as our target approaches for this project.

XLNet is a generalized autoregressive pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and overcomes the limitations of BERT [1] through its autoregressive formulation. XLNet[2] takes inspiration from Transformer - XL into pretraining to give us better performance than BERT in certain scenarios such as sentiment analysis and natural language inference. We believe our use case is one such scenario where we can achieve an improvement with this model. Our plan is to use transfer learning on a pre-trained XLNet model with classification layers on top to achieve our result.

ColBERT is a novel approach suggested by Issa Annamoradnejad et al. [3] for humor detection in short sentences using BERT to generate sentence embeddings and whole text embeddings which are then passed into the neural network to detect existing relationships between sentences and also word level connections in the input text. ColBERT performs better than BERT and XLNet based models for the humor detection task. Based on the approach used, we believe we can reuse this to tackle our target use case as ColBERT is better theoretically at understanding contextual impact of words and the linguistic structure of cyberbullying tweets is similar to how jokes are aligned in a multi-sentence text format.

## 2 Related Work

### Non-word embedding based methods to classify cyberbullying

The first related work[4] in this category highlights the negative effects of cyberbullying on social media on the general populace, before detailing 3 models - Naïve Bayes, Random Forest and J48, they had used to build classify tweets into multiple classes: bully, aggressor, spammer and normal. They also classify the user personalities using Big Five and Dark Triad psychological classifications which are not based in machine learning, to provide as input an estimation of similar historical activity of the author of each tweet. This work presents an interesting non-machine learning based input to add to their classification of tweets, but does not use state-of-the-art methods to perform sentiment analysis on the tweets themselves. Furthermore, these psychological classifications must be performed manually and are expectedly not scalable beyond a few thousand data points. While this paper achieves a high accuracy of around 90.1-91.7% for the different approaches used, we feel our models will perform better as they use more effective models for text classification and are not limited in scalability.

The next related work[5] in this category uses standard deep learning models without word embeddings to perform a binary classification of various sentences into 2 categories - Normal and Insult. The paper implements the following models - Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Units (GRU)[6], Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN), and compares the accuracies between them to find which model performs better. The authors observed that these models achieved accuracies around 81-82% with the best performance achieved using BLSTM. These models struggle compared to the previous related work but inherently address the scalability issue by being completely independent of manual review and user classification. We aim to improve upon this by using word embeddings, which are better suited for text classification as we have learned in class.

The final related work[7] in this category is relatively preliminary paper from 2015 which uses a Naïve Bayes classifier to classify user posts from MySpace.com and Formspring.me into 2 categories - cyberbullying and normal. The paper first tags each word as one of the following classes - noun, pronoun or adjective, and then uses that as input to its model to achieve accuracies around 87-89% in the different websites. This paper, like the others, does not use word embeddings and also focuses on datasets which are currently obsolete. The ways in which bullies target their victims have greatly changed with the improvement in both access and technology in the past 7 years and our work aims to bridge this gap.

### Word embedding based methods to classify cyberbullying

The first related work[8] in this category uses word embeddings - Word2Vec[9], GloVe[10], Reddit[11] and ELMo[12], along with the neural network architectures observed in the second related work in the previous section - GRU, LSTM and BLSTM, in a comparison of the different standard technologies used today. Their experiments are performed on extracted data from the website Formspring.me, which was also referenced in the final related work in the previous section. The paper measures the recall and mean squared error of the different combinations and concludes that the best performance is observed in ELMo with 98% recall when used with BLSTM. We believe this can be improved by using advanced techniques in ColBERT and transfer learning via XLNet which we intend to explore in our work.

The final related work[13] in this category forms the baseline which we intend to compare our models to as they operate on the same dataset we are using for our models - a collection of 47k tweets that have been classified into 6 categories. The paper observes the accuracies and F1 scores of different standard word embeddings - BERT, SBERT, DistilBERT[14], GloVe, Word2Vec, FastText[15], TF-IDF[16], BOW[17], along with some Graph Convolutional Network (GCN) Layers after the embedding layers. The best accuracies and F1 scores for their SOSNet were observed in conjunction with SBERT model at around 92%. We aim to achieve better results than that using our ColBERT and XLNet models.

## 3 Methods

In this section, we will talk about the models we are planning to use for the detection of cyber bullying tweets. The baseline model from reference [13] comprises of Semantic Cosine Similarity Graph Convolutional Network(SOSNet) where there are initial embedding layers(BERT, GloVe, Word2Vec, BOW) and then finally classification layers after the GCN. We use the best performing model observed in the reference paper as the primary baseline model which we compare against.

### ColBERT

We used a pre-trained ColBERT model from reference[3] and fine tuned it to fit the cyberbullying dataset. For preparation of data, we used a tokenizer with a predefined maximum sequence length and a predefined maximum number of sentences an input text can be split into. The individual sentences and the whole text are converted into BERT sentence embeddings before both were passed to input layers of the ColBERT neural network model with parallel hidden layers which amount to 110M parameters. The reasoning behind using both sentence level and whole text sentence embeddings is that, while mid-level features(type of sentence, context) and relations between the sentences are important, it is also useful to obtain word-level connections within the whole text to capture the complete meaning of the sentence. Categorical cross entropy loss will be the loss function used and the final output dense layer will have softmax activation function. We froze all the layers apart from the final 15 dense layers to suit the classification needs.

### XLNet Transfer Learning

Our second approach is to use XLNet and perform multi-class classification via transfer learning. The data is prepared by using the XLNet tokenizer and the input sentences are encoded via the tokenizer. The output dense layers are trainable and will be placed after the frozen XLNet layers and categorical cross entropy loss function is applied over the final dense layer. All of this is done using the SimpleTransformers module to obtain the pre-trained XLNet model.

### Data Pre-processing

Typically, a tweet contains more than just words which can be processed by our model - hashtags, special characters, usernames, images, emojis are all things that interfere with word embedding based models, while simultaneously adding little to no information to our evaluation whether the tweet is cyberbullying and if so, what category. Before we input the tweets into our models, we pre-process our data by pruning this noise to get clean data that contains only the core part of the tweet which will affect our classification of the category of cyberbullying. Word stemming has also been performed.

## 4 Experimental Design

For the purposes of this experiment, the data was split into 80-20 train-test split and pre-processed before being evaluated on the ColBERT and XLNet models. For ColBERT, the input data was tokenized with a maximum sequence length of 20 and with 5 being the maximum sentences for input data to be split into. The sentences and whole text were converted into BERT sentence embeddings before being passed into the parallel input layers, followed by the Custom BERT layer and finally the dense layers to achieve multi-class classification. Categorical cross entropy loss was used and we trained for 150 epochs with a batch size of 64. The optimizer used was Adam with default learning rates.

For XLNet, the input data was encoded via the XLNet tokenizer before being passed to the frozen XLNet layers followed by the trainable output dense layers. Categorical cross entropy loss function was applied over the final dense layer and the training batch size was 32, adam optimizer with default learning rates and decay were used and we trained for 10 epochs.

The baseline model consists of a graph convolutional network approach and different word embedding approaches such as Word2Vec, Glove and BERT are explored. Our results from the ColBERT and XLNet transfer methods are compared with the baseline model as the experiment has been conducted with the same dataset and the results are the best among the reference papers we've come across. The SOSNet approach mainly doesn't use XLNet and ColBERT and this sets up an interesting premise for comparing results and checking if the baseline evaluation scores can be matched or beaten.

For the XLNet model, we experimented with the concept of early stopping where we allow the model to stop training after a certain patience - expressed as a number of epochs, if it observes that a particular metric has remained constant or is not positively changing. This metric could be Matthew's correlation coefficient (MCC) which is a result based on the confusion matrix, or training loss, or training accuracy. We preferred MCC for our evaluation and we tested early stopping with a delta value of 0.01. We also tested the evaluation of the early stopping during the training of the model - after a certain number of training steps as opposed to a certain number of epochs.

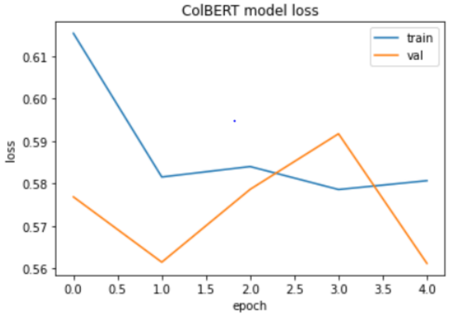
Our evaluation metric is primarily the F1 score. F1 score is an important metric and we use it as the primary evaluation metric as it is the harmonic mean of precision and recall and it sums up the predictive performance of a model by fusing two competing metrics. It is also the primary evaluation metric for the baseline model in their paper. Additionally, we also observed the confusion matrices to evaluate how the model identified each classification category of cyberbullying. We also closely observed the loss curve for the finetuned ColBERT model for the final few epochs that we obtained to gain an insight into the intricacies of that model as it is much more complex and experimental in its design, compared to the XLNet model. Training was done over Google Colab and windows machine with CUDA enabled Nvidia GTX 3060.

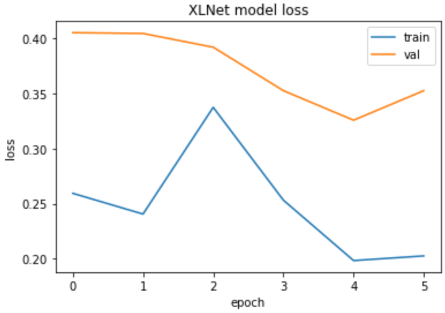
## 5 Experimental Results

Below are the raw results from the two models we've implemented compared to the baseline model, SOSNet + SBERT:

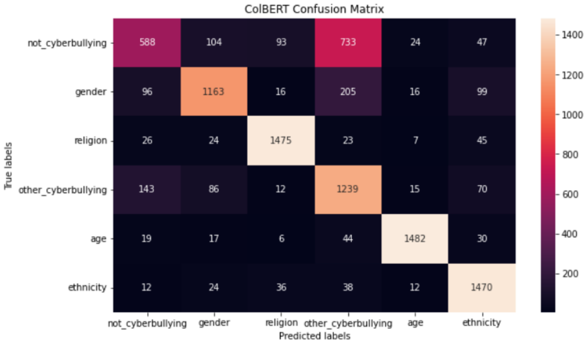
| Metrics | | | |
|---|---|---|---|
| | Precision | Recall | F-1 Score |
| XLNet | 0.86 | 0.86 | 0.86 |
| ColBERT | 0.79 | 0.78 | 0.77 |
| Baseline | 0.92 | 0.92 | 0.92 |

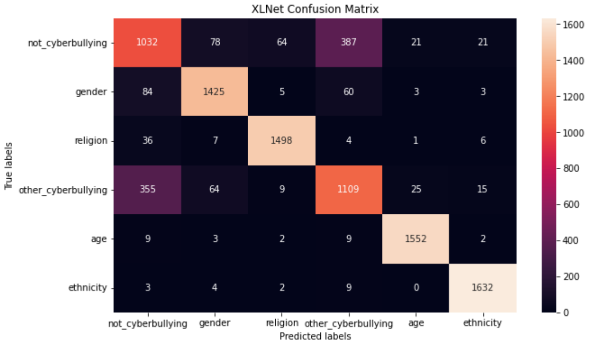
The ColBERT model is trained for 150 epochs with the last 15 dense layers trainable and performed best with data when pre-processed especially with stemming. The XLNet model provided the best results when configured to stop early on MCC metric with a patience of 3 and the model subsequently ran for 6 epochs. We also retained the data pre-processing from ColBERT here and beneath are the loss curves for both the models. The loss curve for the ColBERT model is just of the last 5 epochs as computation was expensive and epochs were run over the span of 1 week with model weights and state being saved periodically. The loss curve for the XLNet model stops after 6 epochs because of early stopping to prevent overfitting as the model observed that the MCC value and subsequently the validation loss were becoming worse.





Shown below are the confusion matrices showing the class-wise distribution of test results over the multi-class classification. There is a clear imbalance against the 'not_cyberbullying' and 'other_cyberbullying' classes where both our models perform poorly with F-1 scores being 0.48 and 0.64 respectively for ColBERT model and 0.66 and 0.70 respectively for the XLNet model for those two classes. The other classes are performing well having F-1 scores ranging from 0.80 to 0.94 for the ColBERT model and 0.90 and 0.98 for the XLNet model.





Initial insights for both models are that the 'gender', 'religion', 'age' and 'ethnicity' classes have well defined linguistic structure and the training data of the dataset conforms to this pattern. Both our models are clearly performing well on these classes and with a bit more refinement with regards to minor model changes and hyperparameter tuning can improve the results even further. On the other hand, 'not_cyberbulling' and 'other_cyberbullying' classes seem to have a relatively lax linguistic structure which is making it harder for the models to pick up vital information and key patterns for multi-class classification.

For the ColBERT model, reducing the number of sentences that a whole text input can be split into and altering the graphical nature of the ColBERT model prior to custom BERT layer, could potentially improve the metrics and pick up the semantic structure of the two problematic classes better. For the XLNet model, adding more dense layers on top of the XLNet structure could potentially pick up the finer details of the classes and improve multi-class classification. However, this must be done while keeping the model from overfitting, which can be assisted by dropout layers.

Experiments with regards to making less or more dense layers of the ColBERT model trainable have yielded roughly similar results for 'not_cyberbulling' and 'other_cyberbullying' classes indicating that the layers prior to custom BERT layer are the ones where the next natural phase of research can go into. As stated, XLNet can be further improved within the model itself for the classes that are already being effectively classified, but this will only be minimal as the model is very optimized. Instead, additional identifying information related to the 'not_cyberbulling' and 'other_cyberbullying' classes would greatly improve the performance of this model for these two classes. Altering the current class balance of the dataset to include more training samples for these two classes may also prove to be fruitful. Techniques of data augmentation could also be explored, such as gaining user information from a model that tracks and tags users based on past user activity.

## 6 Conclusions

From the experiments and the setup of our ColBERT model, we can conclude that cyberbullying does show semantic relations to short humor sentences and our fine-tuned ColBERT model is largely classifying cyberbullying tweets effectively. The intent of cyberbuylling from perpestive of the bully shows relations to how a person may view humor. Also, XLNet is an effective approach for cyberbullying tweet classification into multiple classes, especially linguistically well-structured classes. Our XLNet model currently achieves the best performance metric of the models we implemented, with a F-1 score of 0.86 exceeding or matching some of the models explored in prior literature. It falls short of the baseline SOSNet model in terms of overall metrics but performs better than it in the aforementioned linguistically well-structured classes.

While these models in their current state are acceptable for the purposes of analysis, the logical next step would be to apply these models to flag, report, or remove harmful tweets. Before this can be done, the model must ensure that there are very few false positives as they may lead to censorship of legitimate online discourse or even banning users from social media sites, which have become daily requirements for many people. The models must also constantly evolve to account for the changing English language to be able to keep up with new forms of harassment. Additionally, the dataset must be vetted to account for global geographical, cultural, and ethnic differences and not just be localized to any particular region.

# References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
2. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019)
3. Annamoradnejad, I., Zoghi, G.: Colbert: Using bert sentence embedding for humor detection. arXiv preprint arXiv:2004.12765 (2020)
4. Balakrishnan, V., Khan, S., Arabnia, H.R.: Improving cyberbullying detection using twitter users' psychological features and machine learning. Computers & Security **90** (2020) 101710
5. Iwendi, C., Srivastava, G., Khan, S., Maddikunta, P.K.R.: Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems (2020) 1–14
6. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
7. Nandhini, B.S., Sheeba, J.: Cyberbullying detection and classification using information retrieval algorithm. In: Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015). (2015) 1–5
8. Al-Hashedi, M., Soon, L.K., Goh, H.N.: Cyberbullying detection using deep learning and word embeddings: an empirical study. In: Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems. (2019) 17–21
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
10. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). (2014) 1532–1543
11. Bin Abdur Rakib, T., Soon, L.K.: Using the reddit corpus for cyberbully detection. In: Asian conference on intelligent information and database systems, Springer (2018) 180–189
12. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. corr abs/1802.05365 (2018). arXiv preprint arXiv:1802.05365 (1802)
13. Wang, J., Fu, K., Lu, C.T.: Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In: 2020 IEEE International Conference on Big Data (Big Data), IEEE (2020) 1699–1708
14. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
15. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5** (2017) 135–146
16. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. Volume 242., Citeseer (2003) 29–48
17. Zhang, Y., Jin, R., Zhou, Z.H.: Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics **1**(1) (2010) 43–52