Proceedings of LATERAISSE Workshop @LREC2022, pages 26–31
Marseille, 25 June 2022

Benchmarking Language Models for Cyberbullying Identification and Classification from Social-media texts

Kanishk Verma[a], Tijana Milosevic[b], Keith Cortis[c], Brian Davis[d]
ADAPT SFI Research Centre, Dublin City University[a-d]
DCU Anti Bullying Centre [a-b]
Dublin, Ireland
{kanishk.verma, keith.cortis, brian.davis}@adaptcentre.ie
tijana.milosevic@dcu.ie

Abstract

Cyberbullying is bullying perpetrated via the medium of modern communication technologies like social media networks and gaming platforms. Unfortunately, most existing datasets focusing on cyberbullying detection or classification are i) limited in number ii) usually targeted to one specific online social networking (OSN) platform, or iii) often contain low-quality annotations. In this study, we fine-tune and benchmark state of the art neural transformers for the binary classification of cyberbullying in social media texts, which is of high value to Natural Language Processing (NLP) researchers and computational social scientists. Furthermore, this work represents the first step toward building neural language models for cross OSN platform cyberbullying classification to make them as OSN platform agnostic as possible.

Keywords: Benchmarking, Cyberbullying, Cross-platform, Classification, Transformers

1. Introduction

The *cyberbullying* nomenclature and its propagation medium has evolved over the years, but it can still be understood as a hostile and aggressive behaviour to intentionally and repeatedly hurt or embarrass someone over the internet. Cyberbullying has only exacerbated over recent months due to the COVID-19 pandemic, which has resulted in a surge in online activity among young people. (McBride, 2021), (Raisbeck, 2020), (Jain et al., 2020). Victims of such an act of bullying propagated over the internet may experience lower self-esteem, increased suicidal ideation, and mixed negative emotional responses. (Hinduja and Patchin, 2014).

Recent studies by (Chen and Li, 2020) (Salawu et al., 2020) have leveraged deep neural network (DNN) and neural language modelling (LM) approaches like **B**i-directional **E**ncoder **R**epresentations for **T**ransformers (BERT) by (Devlin et al., 2018) to model cyberbullying detection and classification. As studied by (Emmery et al., 2021), many previous studies in this field of cyberbullying detection are bound by scanty datasets from specific OSN platforms. .

This study aims to develop a cyberbullying text classification language model by evaluating it across multiple Online Social Networking (OSN) platforms to achieve an OSN agnostic cyberbullying classification language model. To that end, we conduct experiments to benchmark pre-trained language models - BERT by (Devlin et al., 2018) and HateBERT by (Caselli et al., 2020) on real-life cyberbullying textual datasets. Although our intent is to cover all OSN platforms, due to the limited nature of the existing research, we are only able to leverage **390,934** sentences or phrases from real-life cyberbullying textual datasets provided by (Hosseinmardi et al., 2015), (Rafiq et al., 2015), (Xu et al., 2012), (Salawu et al., 2020), and (Van Hee et al., 2018). We also establish baselines using traditional Machine Learning (ML) algorithms to benchmark the neural language models.

2. Related Work

Most of the current work in this field by (Tomkins et al., 2018),(Van Hee et al., 2018) ,(Talpur BA, 2020) is focused on social-context-based approaches for binary classification of cyberbullying texts, and these studies rely on Word2Vec by (Goldberg and Levy, 2014), Glove by (Pennington et al., 2014), and FastText (AI, 2015) based word representation techniques. Despite the satisfactory results of recent studies with an amalgamation of NLP and DNN techniques, studies by (Van Hee et al., 2018), (Samghabadi et al., 2020), (Emmery et al., 2019) are bound to ASK.fm data. Studies (Salawu et al., 2020), (Tahmasbi and Rastegari, 2018), (Chatzakou et al., 2017) are restricted to only Twitter data, and studies by (Chen and Li, 2020), (Sourodip Ghosh, 2020), (Paul, 2020) take a multi-modal approach, i.e., text supplemented by social network analysis (SNA)[1] features, are bound to only Instagram and Vine datasets published by (Hosseinmardi et al., 2015) and (Rafiq et al., 2016) respectively.

Other studies by (Sprugnoli et al., 2018), (Bretschneider and Peters, 2016) and a dataset published by (Van Hee et al., 2018) have participant-level annotations that help identify roles of cyberbullying like *harasser, bystander or victim*. Given that the scope of this study focuses on the binary classification of cyberbullying texts, these datasets are not explored for multi-class

[1]SNA: The process of investigating social structures through the use of networks and graph theory

cyberbullying classification, and labels of the dataset by (Van Hee et al., 2018) are converted to binary form, i.e., *bullying* or *non-bullying*.
Also, studies by (Rafiq et al., 2016), (Noviantho et al., 2017), (Al-Ajlan and Ykhlef, 2018), (Hamiza Wan Ali et al., 2018) in cyberbullying text classification have used the traditional ML algorithm Support Vector Machines (SVM) (Wang et al., 2006), as a ML baseline and some other studies by (Dadvar and Eckert, 2018), (Paul, 2020), (Sourodip Ghosh, 2020) have used Bi-directional Long Short-Term Memory (Bi-LSTM) (Huang et al., 2015) for language modelling. To that effect, this study makes the following key contributions,

- First steps to benchmark transformer-based models, neural network and machine learning models for binary classification of cyberbullying texts sourced from real-life cyberbullying textual datasets.
- First steps towards developing an OSN agnostic cyberbullying detection model by training language models on text from one type of OSN platform and evaluating it across multiple OSN platform-types.

3. Experimental Setup

3.1. Datasets

Instagram (IG)[2] dataset sourced from (Hosseinmardi et al., 2015), Vine[3] dataset sourced from (Rafiq et al., 2015), hereafter referred to as *User-Comment datasets* (UC), are similar multimedia content sharing platforms, as they allow users to comment, like and share, multi-media content with one another. ASK.fm[4] and Formspring.me (F.me)[5] datasets sourced from (Van Hee et al., 2018), hereafter referred to as *Question-Answering datasets* (QA) are an anonymous question and answering social networking platform. Twitter[6] datasets sourced from (Xu et al., 2012) and (Salawu et al., 2020), hereafter referred as *Twitter datasets*, are from the OSN platform, Twitter - that allows users to share 280 characters of text as messages termed *tweets*. The lenghts of tokens (words) in each phrase or comment within each of the *seven* dataset, depicts the platform similarity, as represented in the Figure 1. This helps understand that similar platforms have almost similar lengths of tokens. The label and sentence-level details is depicted in Table 1. Each merged dataset is split into 70% for training, 20% for validation, and the remaining 10% is held out for testset. All language models trained on the three merged datasets were evaluated individually across the 10% hold-out test for all merged datasets.

Dataset	Platforms	# of Sentences	Bullying %
User - Comments (UC)	IG + Vine	249,123	23.53%
Question - Answering (QA)	Ask.fm + F.me	129,501	4.60%
Twitter	Twitter	12,310	6.51

Table 1: Percentage-wise Bullying Label Distribution and Sentence Count of all datasets

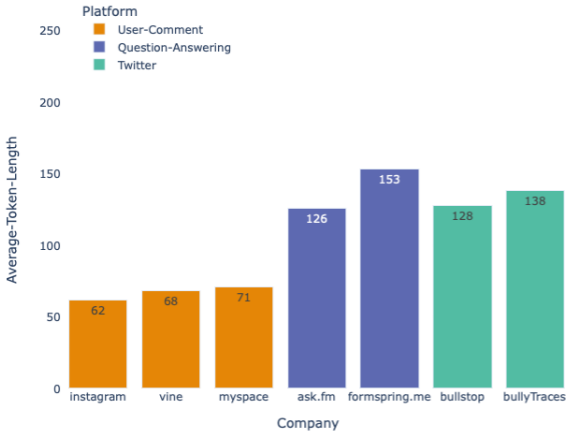

Figure 1: Average lengths of tokens in datasets

3.2. Data Imbalance

There is a high imbalance skewed toward the non-bullying class in all datasets, as depicted in Figure 2. Handling the imbalance is paramount to avoid any learning bias towards the majority class. As the dataset is limited in bullying instances, to avoid any risk of contextual loss and not to alter the sequence of words in sentences, we ruled out the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) and the random under-sampling technique (Prusa et al., 2015). Instead, we leveraged the random over-sampling technique (Fernández et al., 2018), i.e., a technique that duplicates examples of minority class randomly, to balance the data towards the majority class.

3.3. Data Pre-processing

Adhering to General Data Protection Regulation (GDPR) directive (Council of European Union, 2016), we fully anonymised and normalised the datasets for any PII[7] data by leveraging GATE Cloud (Tablan et al.,

[2] https://about.instagram.com/
[3] https://en.wikipedia.org/wiki/Vine_(service)
[4] https://ask.fm/
[5] https://en.wikipedia.org/wiki/Spring.me
[6] https://about.twitter.com/en
[7] Refers to Personally identifiable information

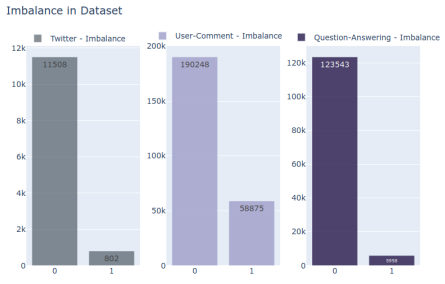

Figure 2: Imbalance in Dataset

2013). Furthermore, the TwitIE API (K. Bontcheva, 2013) to extract named entities in the text. In addition, we also a) removed URLs, user mentions, and non-ASCII characters for all datasets, b) retweet (RT) markers in text for *twitter* datasets, c) lower-cased all text, and d) converted textual contractions to formal format.

3.4. Language Models

Each neural classifier is fine-tuned by adding a fully connected layer on top of its respective pre-trained model.

- **BERT**: Provided by (Devlin et al., 2018), with 12 layers, also known as *transformer blocks*, and trained with 110 M parameters. We fine-tune pre-trained $BERT_{base-uncased}$ language model with different hyperparamters. (See Section 3.5).

- **HateBERT**: Provided by (Caselli et al., 2020), it is a re-trained $BERT_{base-uncased}$ language model, trained on comments from RAL-E Reddit's banned communities[8]. Further pre-training of BERT model is an effective and cost ineffective strategy to port pre-trained language model for other language specific tasks.

- **Bi-LSTM**: A baseline deep neural network model based on Bi-directional Long Short-Term Memory (Huang et al., 2015). We trained this model on five epochs with different hyper-parameters and pre-trained 50 dimensional GloVe-based Twitter word-embedding.

- **Support Vector Machines (SVM)**: A traditional machine learning algorithm known as Support Vector Machines proposed by (Hearst et al., 1998). This algorithm is trained by leveraging Term Frequency - Inverse Document Frequency (TF-IDF) (Zhang et al., 2011).

[8] `https://en.wikipedia.org/wiki/Controversial_Reddit_communities`

3.5. Hyper-parameters and evaluation

- **Hate-BERT** & **BERT**: Our experiments[9], utilised the implementations provided by HuggingFace's Transformer library (Wolf et al., 2019) and the authors of HateBERT. We used the *ModelForSequenceClassifcation* which matches BERT model to the proper implementation. We trained the transformer-based models for `2,3,4` epochs and fine-tuned each model for all 3 merged datasets individually and thus the maximum sequence length varied between `128` to `256` tokens depending on the dataset. We fine-tuned the classification layer for transformer-based models using *ReLU* and the *Adam Weighted* optimizer by (Kingma and Ba, 2015) with a learning rate ranging from `0.1`, `0.001`, $1e^{-5}$ to $5e^{-5}$.

- **Bi-LSTM**: For the Bi-LSTM model, the recurrent dropout rate was set to 0.2 and the fully connected layer was added with 256 neurons and *ReLU* activation function, and since it was engineered for a binary task, the output layer was set to *softmax*. The *Cross Entropy loss* function was used for fine-tuning both transformer-based models and training Bi-LSTM.

- **SVM**: For the SVM model, we first conducted a grid search with five cross-validation and the hyper-parameters from the best model were used for training.

To benchmark and evaluate the fine-tuned transformer-based models, we conducted experiments with one traditional approach using SVM with TF-IDF and one Bi-LSTM algorithm with GloVe-based pre-trained 50-dimensional vectors. In addition, we evaluate the performance of these language models based on F1 scores (Chinchor and Sundheim, 1993) for positive (bullying) and negative (non-bullying) and overall F1 scores. F1 scores consider both the precision and recall to compute their metrics, and it can be interpreted as the weighted average of the two classes.

4. Results

As indicated in Table 2, the fine-tuned Hate-BERT language model has a significant advantage over the fine-tuned BERT, Bi-LSTM and traditional SVM. Although our experiment results indicated in the Table 2 show that models trained and tested on texts from the same OSN platform perform better when evaluated across different OSN platforms. The Hate-BERT language model, when fine-tuned on the *Question-Answering* datasets (ASK.fm and Formspring.me) and *Twitter* datasets for binary classification of cyberbullying text, has outperformed other baselines earlier discussed in the Section 3.4. Although the SVM model

[9] All the experiments in this work were conducted on a local system with a 16 core CPU, 16GB RAM and a NVIDIA RTX 2070 GPU (8GB GPU Memory)

trained on *user-comment* datasets (Myspace, Vine, Instagram) performs well with a **0.75** F1 score in classifying bullying samples as bullying, the same model only performs with **0.56** F1-score for classifying bullying samples. The Bi-LSTM model trained on *Twitter* datasets performs well with a **0.69** F1-score for classifying bullying samples, the same model achieves **0.63** F1-score for classifying bullying samples for the *user-comment* dataset. Additionally, our experiments depict that when the Hate-BERT model is fine-tuned on the *Question-Answering* datasets, it is able to achieve **0.73** F1-score in classifying bullying samples for both *user-comment* and *question-answering* dataset. Moreover, when we fine-tune the Hate-BERT model on *twitter* datasets, though it achieves **0.78** F1-score for *twitter* datasets, it is only able to achieve **0.71** F1-score for classifying bullying samples for the *user-comment* dataset. These exhaustive experiments indicate that fine-tuning language models from three OSN platforms are the first step toward developing an OSN platform-agnostic cyberbullying detection mechanism. Moreover, our results suggest that more work will be beneficial in developing such platform-agnostic detection mechanisms.

Model	**Train-set**	**Test-set**	**Bully F-1**	**Non-bully F1**	**Avg F1**
SVM	UC	UC	**0.75**	0.71	**0.73**
		QA	**0.56**	0.58	**0.57**
		Twitter	0.51	0.51	0.51
	QA	UC	0.34	0.50	0.42
		QA	0.52	0.54	0.53
		Twitter	0.52	0.52	0.52
	Twitter	UC	0.28	0.50	0.39
		QA	0.48	0.50	0.49
		Twitter	0.54	0.54	0.54
Bi-LSTM	UC	UC	0.68	0.70	0.69
		QA	0.30	0.50	0.40
		Twitter	0.51	0.51	0.51
	QA	UC	0.58	0.60	0.59
		QA	0.69	0.67	0.68
		Twitter	0.52	0.54	0.53
	Twitter	UC	**0.63**	0.61	**0.62**
		QA	0.61	0.57	0.59
		Twitter	**0.69**	0.67	**0.68**
BERT	UC	UC	0.65	0.77	0.71
		QA	0.54	0.58	0.56
		Twitter	0.58	0.60	0.59
	QA	UC	0.48	0.50	0.49
		QA	0.62	0.68	0.65
		Twitter	0.57	0.61	0.59
	Twitter	UC	**0.54**	0.52	**0.53**
		QA	0.63	0.61	0.62
		Twitter	**0.75**	0.79	**0.77**
Hate-BERT	UC	UC	**0.68**	**0.84**	**0.76**
		QA	0.67	0.59	0.63
		Twitter	**0.65**	0.71	**0.68**
	QA	UC	**0.73**	0.65	**0.69**
		QA	**0.73**	0.73	**0.73**
		Twitter	0.61	0.65	0.63
	Twitter	UC	**0.71**	0.65	**0.68**
		QA	0.69	0.65	0.67
		Twitter	**0.78**	0.84	**0.81**

Table 2: All Results

5. Conclusion & Future Work

We have provided a comprehensive benchmark on the binary classification of cyberbullying in a social media text. Our experiments demonstrate that merging existing datasets from similar platforms can improve the performance of transformer-based models. Also, fine-tuning the pre-trained Hate-BERT model outperforms the BERT, Bi-LSTM and SVM models. This novel benchmarking study is the first step toward building an OSN agnostic neural language model for the cyberbullying domain. One limitation of our study is that we use word-count (TF-IDF) and non-contextual word-embeddings (Glove) for text representation while training the baseline models - SVM and Bi-LSTM. Instead, future research should leverage contextual word embeddings from BERT and Hate-BERT language models for training these baseline models. The current availability of datasets and resources in the area of cyberbullying, as highlighted by (Emmery et al., 2021) and observed in this study, is scarce and highly skewed to negative class, i.e., to non-bullying instances. Therefore, there is a need to divulge qualitative and not quantitative cyberbullying research to build better language models to detect cyberbullying. Moreover, a detailed ablation study of the language models will aid in future benchmarking of such cyberbullying classifiers. In addition, it will help clarify how language models better classify specific samples from certain classes than the others.

6. Acknowledgements

We would like to thank the authors (Hosseinmardi et al., 2015), (Rafiq et al., 2015), (Van Hee et al., 2018), (Xu et al., 2012), (Salawu et al., 2020) for sharing the data.

This research has received funding from the *Irish Research Council* and *Google* under grant number EPSPG/2021/161, *Facebook/Meta Content Policy Award*, Phase 2: Co-designing with children: A rights-based approach to fighting bullying. In addition, this research has also received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801522, by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology grant number 13/RC/2106_P2.

7. Bibliographical References

AI, F. (2015). Fasttext. https://ai.facebook.com/tools/fasttext/, November. Facebook AI.

Al-Ajlan, M. A. and Ykhlef, M. (2018). Optimized twitter cyberbullying detection based on deep learning. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–5.

Bretschneider, U. and Peters, R. (2016). Detecting cyberbullying in online communities.

Caselli, T., Basile, V., Mitrovic, J., and Granitzer, M. (2020). Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472.

Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E. D., Stringhini, G., and Vakali, A. (2017). Mean birds: Detecting aggression and bullying on twitter. *CoRR*, abs/1702.06877.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, H.-Y. and Li, C.-T. (2020). Henin: Learning heterogeneous neural interaction networks for explainable cyberbullying detection on social media. *arXiv preprint arXiv:2010.04576*.

Chinchor, N. and Sundheim, B. M. (1993). Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Council of European Union. (2016). Regulation (eu) 2016/679 of the european parliament and of the council (general data protection regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Dadvar, M. and Eckert, K. (2018). Cyberbullying detection in social networks using deep learning based models; A reproducibility study. *CoRR*, abs/1812.08046.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Emmery, C., Verhoeven, B., Pauw, G. D., Jacobs, G., Hee, C. V., Lefever, E., Desmet, B., Hoste, V., and Daelemans, W. (2019). Current limitations in cyberbullying detection: on evaluation criteria, reproducibility, and data scarcity. *CoRR*, abs/1910.11922.

Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., Desmet, B., Hoste, V., and Daelemans, W. (2021). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. *Language Resources and Evaluation*, 55(3):597–633.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced data sets*, volume 10. Springer.

Goldberg, Y. and Levy, O. (2014). word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

Hamiza Wan Ali, W. N., Mohd, M., and Fauzi, F. (2018). Cyberbullying detection: An overview. In *2018 Cyber Resilience Conference (CRC)*, pages 1–3.

Hearst, M., Dumais, S., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Hinduja, S. and Patchin, J. W. (2014). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin press.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, pages 49–66. Springer.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Jain, O., Gupta, M., Satam, S., and Panda, S. (2020). Has the covid-19 pandemic affected the susceptibility to cyberbullying in india? *Computers in Human Behavior Reports*, 2:100029.

K. Bontcheva, L. Derczynski, A. F. M. G. D. M. N. A. (2013). witie: An open-source information extraction pipeline for microblog text. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

McBride, M. (2021). Cyberbullying soared during lockdown. what are schools doing about it? https://www.irishtimes.com/news/education/cyberbullying-soared-during-lockdown-what-are-schools-doing-about-it-1.4473011, Feb. The Irish Times.

Noviantho, Isa, S. M., and Ashianti, L. (2017). Cyberbullying classification using text mining. In *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*, pages 241–246.

Paul, S., S. S. (2020). Cyberbert: Bert for cyberbullying identification. *Multimedia Systems*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., and Napolitano, A. (2015). Using random undersampling to alleviate class imbalance on tweet sentiment data. In *2015 IEEE International Conference on Information Reuse and Integration*, pages 197–202.

Rafiq, R. I., Hosseinmardi, H., Han, R., Lv, Q., Mishra,

S., and Mattson, S. A. (2015). Careful what you share in six seconds: Detecting cyberbullying instances in vine. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 617–622. IEEE.

Rafiq, R. I., Hosseinmardi, H., Mattson, S. A., Han, R., Lv, Q., and Mishra, S. (2016). Analysis and detection of labeled cyberbullying instances in vine, a video-based social network. *Social network analysis and mining*, 6(1):1–16.

Raisbeck, D. (2020). Experts around the world warn parents to be vigilant as cyberbullying increases during lockdown. `https://www.cybersmile.org/news/experts-around-the-world-warn-parents-to-be-vigilant-as-cyberbullying-increases-during-lockdown`. The Cybersmile Foundation.

Salawu, S., He, Y., and Lumsden, J. (2020). Bullstop: A mobile app for cyberbullying prevention. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 70–74.

Samghabadi, N. S., Monroy, A. P. L., and Solorio, T. (2020). Detecting early signs of cyberbullying in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 144–149.

Sourodip Ghosh, Aunkit Chaki, A. K. (2020). Cyberbully detection using 1d-cnn and lstm. *Proceedings of International Conference on Communication, Circuits and Systems*.

Sprugnoli, R., Menini, S., Tonelli, S., Oncini, F., and Piras, E. (2018). Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium, October. Association for Computational Linguistics.

Tablan, V., Roberts, I., Cunningham, H., and Bontcheva, K. (2013). Gatecloud.net: a platform for large-scale, open-source text processing on the cloud. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1983):20120071.

Tahmasbi, N. and Rastegari, E. (2018). A socio-contextual approach in automated detection of public cyberbullying on twitter. *Trans. Soc. Comput.*, 1(4), December.

Talpur BA, O. D. (2020). Cyberbullying severity detection: A machine learning approach.

Tomkins, S., Getoor, L., Chen, Y., and Zhang, Y. (2018). A socio-linguistic model for cyberbullying detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 53–60.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.

Wang, Z.-q., Sun, X., Zhang, D.-x., and Li, X. (2006). An optimal svm-based text classification algorithm. In *2006 International Conference on Machine Learning and Cybernetics*, pages 1378–1381.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666.

Zhang, W., Yoshida, T., and Tang, X. (2011). A comparative study of tf*idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765.