Identification and Classification of Cyberbullying Posts: A Recurrent Neural Network Approach using Under-sampling and Class Weighting

Ayush Agarwal[1], Aneesh Sreevallabh Chivukula[2], Monowar H. Bhuyan[3], Tony Jan[4], Bhuva Narayan[5] and Mukesh Prasad[2]

[1] Department of Information Technology, Delhi Technological University, India
[2] School of Computer Science, FEIT, University of Technology Sydney, Australia
[3] School of IT and Engineering, Melbourne Institute of Technology, Australia
[4] Department of Computing Science, Umea University, Sweden
[5] School of Communication, FASS, University of Technology Sydney, Australia

Abstract. With the number of users of social media and web platforms increasing day-by-day in recent years, cyberbullying has become a ubiquitous problem on the internet. Controlling and moderating these social media platforms manually for online abuse and cyberbullying has become a very challenging task. This paper proposes a Recurrent Neural Network (RNN) based approach for the identification and classification of cyberbullying posts. In highly imbalanced input data, a Tomek Links approach does under-sampling to reduce the data imbalance and remove ambiguities in class labelling. Further, the proposed classification model uses Max-Pooling in combination with Bi-directional Long Short-Term Memory (LSTM) network and attention layers. The proposed model is evaluated using Wikipedia datasets to establish the effectiveness of identifying and classifying cyberbullying posts. The extensive experimental results show that our approach performs well in comparison to competing approaches in terms of precision, recall, with F1 score as 0.89, 0.86 and 0.88, respectively.

Keywords: Cyberbullying, Natural Language Processing, Under-sampling, Recurrent Neural Network, Social Media

1 Introduction

There has been a dramatic increase in instances of online abuse and cyberbullying on web platforms such as Wikipedia, YouTube, Instagram, Reddit, Facebook, and Twitter in the recent years. Being able to comment or reply anonymously has further fuelled the growth of such instances. According to Chu et al. [1], 40% of people on the web have experienced bullying or harassment of some kind including sexual harassment, physical threats, etc. In extreme cases, cyberbullying can cause severe mental health issues as well. Manually filtering the comments or replies that qualify as cyberbullying can be a very tedious, if not an impossible task when there are hundreds of thousands of comments being posted every hour. Hence, there has been an increasing demand for developing ways to detect instances of cyberbullying automatically and filter them out without human intervention.

Due to increasing availability of annotated datasets from web platforms (e.g., Facebook, YouTube), we can leverage machine learning and natural language processing techniques for data-driven solutions to detect cyberbullying posts. Deep learning has also evolved as an efficient solution for such cyberbullying detection problems due to the availability of a large amount of labeled data for supervised learning. However, building highly accurate models for cyberbullying detection remains difficult for several reasons. As outlined by Wulczyn et al. [2], firstly, even though there is a definition for cyberbullying, there are no hard guidelines to determine if a piece of text may constitute a cyberbullying comment or not. This can often be highly dependent on the context of the comment. As observed in crowdsourced datasets, not every annotator has the same opinion about each comment, and the annotations are dependent on the annotator's bias. Secondly, publicly-available datasets are highly imbalanced and have a very small percentage of comments labeled as positive for bullying. Machackova et al. [23] discuss attack patterns of cyberbullying and coping strategies used by different groups. They measure the effect of cyberbullying in terms of the type of attack, length of cyber aggression, harm experienced by the person and how the user responds to the attack. Using crowdsourcing, Wulczyn et al. [2] released cyberbullying datasets over a large corpus of over 100k human-annotated comments on Wikipedia articles. The corpus is annotated to indicate if a comment indicates a personal attack. They have also performed a thorough analysis of the data to answer questions related to anonymity and patterns of attacks. This paper leverages the Wikipedia datasets in modelling evaluation.

Improving upon the approaches using Term Frequency-Inverse Document Frequency (TFIDF), Yin et al. [3] combined features like context, sentiment, and content for designing a supervised classification model in cyberbullying detection. Tokunga [4] provides a review of past research work on cyberbullying victimization. It discusses the evolving definition of the term Cyberbullying and provides research directions to better theorize the detection problem. Schrock and Boyd [5] list the major platforms where Cyberbullying may take place: chat-rooms, social media websites, blogs, and multiplayer online games. Warner and Hirschberg [6] present an annotated corpus for words that are commonly found in hate speech texts. This approach was to feed feature sets in an SVM classifier. Kwok and Wang [7] implement a binary Naïve Bayes classifier on a Twitter dataset for classifying tweets as racist and not racist. But their model did not achieve significant performance gains. Cheng et al. [8] present antisocial behavior analysis based on online forums and report an analysis of the commenting patterns of people who tend to get banned from these forums due to their behavior. Waseem and Hovy [9] propose a list of eleven conditions for annotating a tweet corpus created by them for studying hate speech. They experiment with variable length character n-grams used for performing a binary classification through logistic regression. Waseem [10] performs an analysis of annotator behavior on the corpus. They found that machine learning systems trained on annotations created by experts rather than amateurs were better at predicting hate speech.

Ross et al. [11] assess the reliability of annotations for detecting hate speech and cyberbullying. They used the definition of hateful comments provided by Twitter to see if it improved annotation quality which in their study did not. They motivate the need for a more pervasive definition of cyberbullying which would guide the annotator's behaviour towards creating more reliable annotations. Nobata et al. [12] propose a supervised approach for online abuse detection by extracting four types of features, namely, syntactic, distributed syntactic, linguistic and n-grams. They further performed a temporal analysis of the data to analyze its robustness.

Saleem et al. [13] bypass the annotation problem entirely by finding online communities that identify themselves as self-hate on Reddit. Their method performed better than the earlier keyword-based approaches even while using logistic regression for performing classification. Sahlgren et al. [14] propose a method for learning textual representations for abusive languages through three approaches, keywords, n-grams, and word embeddings. Their method was tested on the Wikipedia dataset using a logistic regression classifier. Aroyehun and Gelbukh [15] experiment with deep learning models like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) for performing the classification on an aggression dataset. For performance improvement, they augmented the training dataset using round-trip translation. Chu et al. [1] assess the Wikipedia dataset with three different deep learning models, LSTM with word embeddings, CNN with word embeddings and CNN with character embeddings. They used glove vectors for initializing the embedding matrix of training data. Cheng et al. [8] propose another approach of using multi-task sentence embedding models using SVM Classifier for abuse and cyberbullying detection. Mishra et al. [16] derive a method for generating context-aware embeddings for out-of-vocabulary words. They then use bi-directional RNNs for performing the classification and posted their results on the Wikipedia dataset. Kumar et al. [17] employ a concatenated attention and bi-directional RNN model for modeling the semantic and contextual relations in the text.

The following are the major contributions of this paper:

- We propose a RNN-based approach to identify and classify the cyberbullying posts.
- We use the word embeddings from two different sources to initialize the model and uses max-pooling to reduce the sparseness of the data representation in an embedding layer.
- Then we use multiple Bi-LSTM layers along with attention for processing contextual information in the text.
- Finally we perform under-sampling and use class weighting to reduce the effect of class imbalance in the dataset on classification model's training loss and testing performance.

The rest of the paper is organized as follows: Section 2 explains the proposed approach, Section 3 describes the dataset and experimental results, and finally future research directions are given in the conclusion in Section 4.

2 Proposed Approach

The text input embeddings are initialized by performing a mean of 300-dimensional glove [24] vectors and 300-dimensional paragram [25] embedding for each word in the vocabulary. We use two different embeddings to accommodate the vocabulary of the

Wikipedia dataset as 1/3 of the vocabulary of the dataset was not present in the glove word embeddings. Due to the varied length of comments and some uncommon vocabulary on Wikipedia, the embedding matrix was very sparse. After the text embedding layer, we propose a 1-dimensional max-pooling layer to reduce the sparseness of the embedding matrix by reducing the total number of values in the matrix by half (with window size = 2). This ensured that sparseness was reduced while losing minimum information because of the small window size.

Bidirectional LSTMs (Bi-LSTMs) [19] are an extension of traditional LSTMs. Bi-LSTMs train two LSTMs instead of one LSTM on the input sequence. The first LSTM is trained on the input sequence as-is and the second LSTM is trained from the opposite direction on a reversed copy of the input sequence. Hence, it is possible to capture the contextual information in a much better way as the information can be processed from both the previous and future time stamps. As a more efficient choice for understanding sequential information, our proposed architecture uses multiple blocks of bi-directional LSTM layers for capturing contextual features in the comments. The proposed model also contains a hierarchical attention layer for focusing on more important words. An input text may contain a lot of irrelevant words which is not important for classification. Attention mechanisms [20] allow us to attend to or focus on the more relevant words of such an input by giving them a higher importance in classification. With an attention mechanism, the full source sentence isn't encoded into a fixed-length vector. Rather, a decoder network is allowed to "attend" to different parts of the source sentence at each step of the output generation. Importantly, this lets the model learn what to attend to, based on the input sentence and what it has produced so far.

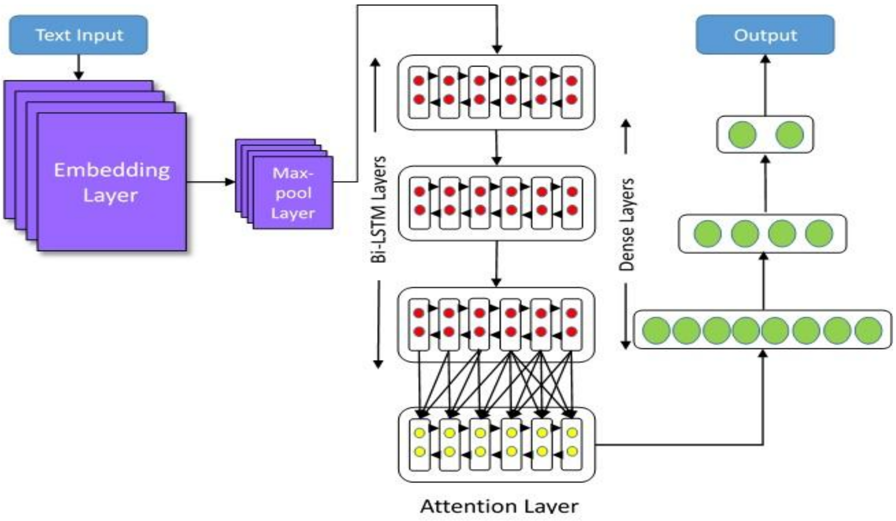

Fig. 1. Architecture of the proposed approach

The proposed architecture of the model is shown in Figure 1. The processed text input is sent to the Embedding Layer. Max pooling is applied to the output received from the Embedding Layer. The pooled output is sent to a stack of 3 Bi-directional

LSTM Networks and an Attention Layer. The output received from this stack is sent to a network of dense layers. The output is then classified using a softmax classifier. Class weighting is applied to counter the imbalance in the data. Samples in each class are given different weights while calculating the training loss. These class weights are inversely proportional to the number of samples in the class. Thus we give more weightage to the minority class calculating the training loss. The loss function is thus penalized more for misclassifying a sample belonging to the minority class.

3 Performance evaluation

3.1 Dataset

To establish the performance of the proposed approach, we use the Wikipedia dataset [2], which contains over 100k comments from the discussions in the talk pages of Wikipedia articles. The comments were labeled using crowdsourcing with 10 annotators. Of the 100k comments, 13,590 have been labeled as a personal attack, and rest as not containing any personal attack. The dataset was cleaned to remove white spaces, special characters, punctuation, digits, contractions and some common misspellings were also corrected. The comments were then tokenized. Table 1 is a summary of the Wikipedia dataset for cyberbullying detection. Figure 3 and Figure 4 display a word cloud of the most frequent keywords in the non-personal attack and personal attack category of the dataset, respectively.

Table 1. Wikipedia Dataset

Dataset	No. of Posts	Max Length of Comments	95 percentile length of comments	No. of Classes	Vocabulary Size
Wikipedia	100,000	2846	231	2	55262

Under-sampling refers to a group of sampling techniques designed to balance the class distribution of an imbalanced dataset. They are generally used to reduce samples from the majority class to reduce or eliminate imbalance in classes in proportion to samples. In contrast, oversampling adds samples to the minority class to reduce the class imbalance. However, over-sampling can often lead to overfitting due to the repetition of samples of the minority class.

In the Wikipedia dataset, the minority class was only a little more than 10% of the dataset. We use Tomek Link under-sampling method [21] to reduce this data imbalance. Tomek Link method is used for removing samples that lie on the borderline of pairs of classes. Given two instances x and y belonging to different classes and separated by a distance dist(x,y), (x,y) is called a Tomek link if there is no instance z such that dist(x,z) is less than dist(x,y) or dist(y,z) is in turn less than dist(x,y). Thus, Tomek Links Under-sampling method removes a sample (A) which satisfies this condition i.e. there is no other sample (B) who's distance from (A) is less than (A)'s distance from origin. Some of the top words from the comments labeled as a not a personal attack and personal attack are shown in Figures 3 and 4, respectively.

standard
three
history
book
cell
definition
even
will
article
re
make
understand
made
insurance
quote
man
destruction
likely
neutral
ensurance
context
someone
fine
term
comes
use
particle
age
assured
creative
cosmology
proposal
mutual
don
peace
speak
think
popper
criticism
false
rule
eye
multiple
beyond
family
admin
describe
legitimate
arab
warning
gandhi
notable
require
want
kari
mitchell
versus
hunter

Fig. 3. Some of the top words from the comments labeled as a not a personal attack

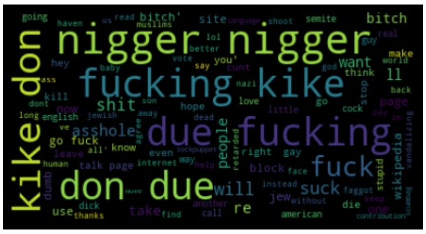

Fig. 4. Some of the top words from the comments labeled as a personal attack

3.2 Results

The dataset is divided into an 80-20 percentage split. The validation data is further divided into test data and validation data according to a 50-50 percentage split. We report the precision, recall and F1 score on the testing data for the proposed RNN model. Training loss is calculated using binary cross-entropy loss function performing the classification with a softmax classifier and an Adam optimizer [22]. Class weighting scheme, is also used to account for the imbalanced data, for further optimization of the training loss. The proposed model achieves 0.89 precision, 0.86 recall and 0.88 F1 score for the test data in Table 2. The performance is compared with the results achieved by Mishra et al. [16], Kumar et al. [17], Chu et al. [1] and Chen et al. [18] in Table 2. The experimental setup for data partition and calculation of precision, recall, F1-score, accuracy and validation approaches for the proposed approach is the same as other approaches with which the result has been compared.

Table 2. Performance comparison of the proposed approach with other methods on the test dataset

Methods	Precision	Recall	F1 score
Mishra et al. [16]	0.81	0.74	0.77
Kumar et al. [17]	0.83	0.77	0.79
Chu et al. [1]	----	----	0.71
Chen et al. [18]	----	0.82	----
The Proposed Approach	**0.89**	**0.86**	**0.88**

In the proposed method, Tomek link under-sampling helps to remove data samples that may be ambiguous or borderline for the training algorithm to correctly classify. As the text length and vocabulary is highly varied, Max-pooling also reduces the sparseness of the embedding matrix. Bi-directional LSTM Layers make it possible to understand the contextual properties of the cyberbullying comments. Attention Layer helps in attending to the most important parts of the text.

4 Conclusion and future work

In this paper, the challenges of detecting cyberbullying in online comments are addressed. The use of under-sampling and class weighting schemes in the training loss function reduces the effect of class imbalance in classifying the dataset. Multiple blocks of bi-directional LSTM and attention layers capture contextual and temporal information in the data. Max-pooling reduces sparseness of the embedding matrix representing cyberbullying text. Proposed modelling elements combine together to increase classification performance in cyberbullying detection. The proposed approach performs significantly better than the approaches in the state-of-the-art ones on Wikipedia datasets. This method can be employed for the automatic detection of online cyberbullying. Next, we will experiment with sparse representation models and deep generative modelling on the embedding matrix, and explore attention mechanisms for imbalanced classification.

References

[1]. T. Chu, K. Jue, and M. Wang, "Comment abuse classification with deep learning," *Von https://web. stanford. edu/class/cs224n/reports/2762092. pdf abgerufen,* 2016.

[2]. E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in Proceedings of the 26th International Conference on World Wide Web, 2017, pp. 1391-1399.

[3]. D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1-7, 2009.

[4]. R. S. Tokunaga, "Following you home from school: A critical review and synthesis of research on cyberbullying victimization," Computers in human behavior, vol. 26, no. 3, pp. 277-287, 2010.

[5]. A. Schrock and D. Boyd, "Problematic youth interaction online: Solicitation, harassment, and cyberbullying," Computer-mediated communication in personal relationships, pp. 368-398, 2011.

[6]. W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proceedings of the second workshop on language in social media, 2012: Association for Computational Linguistics, pp. 19-26.

[7]. I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in Twenty-seventh AAAI conference on artificial intelligence, 2013.

[8]. J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," in Ninth International AAAI Conference on Web and Social Media, 2015.

[9]. Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in Proceedings of the NAACL student research workshop, 2016, pp. 88-93.

[10]. Z. Waseem, "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter," in Proceedings of the first workshop on NLP and computational social science, 2016, pp. 138-142.

[11]. B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the european refugee crisis," arXiv preprint arXiv:1701.08118, 2017.

[12]. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in Proceedings of the 25th international conference on world wide web, 2016, pp. 145-153.

[13]. H. M. Saleem, K. P. Dillon, S. Benesch, and D. Ruths, "A web of hate: Tackling hateful speech in online social spaces," arXiv preprint arXiv:1709.10159, 2017.

[14]. M. Sahlgren, T. Isbister, and F. Olsson, "Learning representations for detecting abusive language," in Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 2018, pp. 115-123.

[15]. S. T. Aroyehun and A. Gelbukh, "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling," in Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 90-97.

[16]. P. Mishra, H. Yannakoudakis, and E. Shutova, "Neural character-based composition models for abuse detection," arXiv preprint arXiv:1809.00378, 2018.

[17]. R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 1-11.

[18]. H. Chen, S. McKeever, and S. J. Delany, "The Use of Deep Learning Distributed Representations in the Identification of Abusive Text," in Proceedings of the International AAAI Conference on Web and Social Media, 2019, vol. 13, no. 01, pp. 125-133.

[19]. M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, 1997.

[20]. A. Vaswani et al., "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998-6008.

[21]. I. Tomek, "Two modifications of CNN," 1976.

[22]. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[23]. H. Machackova, A. Cerna, A. Sevcikova, L. Dedkova, and K. Daneback, "Effectiveness of coping strategies for victims of cyberbullying," Cyberpsychology: Journal of Psychosocial Research on Cyberspace, vol. 7, no. 3, 2013.

[24]. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532-1543.

[25]. J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," arXiv preprint arXiv:1511.08198, 2015.