Cyberbullying Intervention Interface Based on Convolutional Neural Networks

Qianjia Huang
Department of Computer Science
University of Ottawa
qhuan035@uottawa.ca

Jianhong Zhang
Department of Computer Science
University of Ottawa
jzhan410@uottawa.ca

Diana Inkpen
Department of Computer Science
University of Ottawa
Diana.Inkpen@uottawa.ca

David Van Bruwaene
SafeToNet Canada
dvanbruwaene@safetonet.com

Abstract

This paper describes the process of building a cyberbullying intervention interface driven by a machine-learning based text-classification service. We make two main contributions. First, we show that cyberbullying can be identified in real-time before it takes place, with available machine learning and natural language processing tools, in particular convolutional neural networks. Second, we present a mechanism that provides individuals with early feedback about how other people would feel about wording choices in their messages before they are sent out. This interface not only gives a chance for the user to revise the text, but also provides a system-level flagging/intervention in a situation related to cyberbullying.

1 Introduction

Cyberbullying, which can be defined as *'when the Internet, cell phones or other devices are used to send or post text or images intended to hurt or embarrass another person'* (Dinakar et al., 2012), has become a pernicious social problem in recent years. This is also worrying, as multiple studies found that cyberbullying victims often have psychiatric and psychosomatic disorders (Beckman et al., 2012), and a British study found that nearly half of suicides among young people were related to bullying (BBC News[1]). These factors underscore an urgent need to understand, detect, and ultimately reduce the prevalence of cyberbullying.

In contrast to traditional bullying (e.g., school bullying), cyberbullying is not limited to a time and place, which makes cyberbullying potentially more prevalent than traditional bullying. Cyberbullying victims may not recognize their experiences as bullying and they may not report them or seek help for associated emotional difficulties. Kowalski and Limber (2007) reported that almost 90% of young cyberbullying victims did not tell their parents or other trusted adults about their online negative experiences. These factors are especially worrying as multiple studies have reported that the victims of cyberbullying often deal with psychiatric and psychosomatic disorders (Beckman et al., 2012; Sourander et al., 2010), and the worst cases are suicides (Tokunaga, 2010).

Given the importance of the problem, content-based cyberbullying detection is becoming a key area of cyberbullying research. Current state-of-the-art methods for cyberbullying detection

[1] http://www.bbc.co.uk/news/10302550

Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pages 42–51
Santa Fe, USA, August 25, 2018.

combine contextual and sentiment features (e.g., curse word dictionaries, histories of users activities, grammatical properties, and sentiment features derived from online users content) with text-mining approaches. While performance can be improved by training on text-external features, the scarcity of platform-ubiquitous external features requires a cross-platform new-media text classification algorithm to be trained strictly on text. This reduces the presence of features that are significant in training, but absent from data used in out-of-domain contexts. Hence, we introduce a text-driven model which covers six social media platforms (Facebook, Instagram, Twitter, Pinterest, Tumblr, Youtube), and it could be an ideal solution for this problem.

While presenting their methods for cyberbullying detection7, scholars have also suggested different interfaces for intervention. Dinakar et al. (2012) describe cyberbullying intervention mock-ups for both sender and receiver. With the aid of the text-driven model, this project also implements an Android-based interface which combines and optimizes the mock-ups from Dinakar et al. (2012). For instance, an interface giving the sender a chance to retype/cancel the message (as shown in Figure 1) is considered in our project. This project could be developed into a thrid-party application between users and social media providers for creating a healthy online environment.

Figure 1: Mock-up for delaying and undoing the issuance of messages from Dinakar et al. (Dinakar et al., 2012)

2 Online Cyberbullying Model

2.1 Data Set

This paper uses a dataset that was built for cyberbullying detection. Visr, a predictive wellness company, produced this dataset for use with an application (app) that analyzes online activities and interactions, and then alerts parents to potentially harmful issues their children may be experiencing.[2] Issues that parents are alerted about include bullying, anxiety, and depression. By making parents immediately aware of emerging issues on Instagram, Gmail, Tumblr, YouTube, Facebook, Twitter, and Pinterest, the Visr app aims to help parents address such issues before they grow into thornier problems. The app raises a red flag to warn parents when signs of these issues are detected in a child's online activities, including signs of possible mental health consequences like nascent depression, eating disorders, and self-harm. Visr accesses children's social media content through the API's of these social media channels with the consent of the children who are the account holders. This provides a unique cross-platform dataset with rich information.

[2]The app is available at https://app.visr.co, through the Apple App store, or through the Google Play store.

The data was collected by the Visr child safety app from September 2014 to March 2016. Over a half-million online posts were selected from among the social media platforms (Facebook, Instagram, Twitter, Pinterest, Tumblr, Youtube) and Gmail. These posts were randomly chosen among posts that had been viewed, received, or sent by the adolescents (between age 13 to 17). Personally identifying information was removed to ensure the privacy of Visr users. Demographic information such as gender, age, location's time-zone, post time, and the number of likes were recorded as well.

The specific cyberbullying detection dataset is one of Visr's labeled datasets. After combining an enriched selection of those posts with 3,072 'real issues' (posts with issues confirmed by parents), an annotation process was performed by three annotators for the *cyberbullying* label. 1,753 posts were determined to be positive for *cyberbullying*), 304 were labeled as 'unsure', and 12,441 were labeled as negative for *cyberbullying* (Agreement percentage: 95.07%; Cohen's Kappa: 0.805). The posts either labeled as 'unsure' or about which annotators disagreed were removed, leaving a corpus of 14,194.

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are known to have good performance on data with high locality, when words get more care weight about the features surrounding them. For our classification problem, we are trying to get high locality in text given their short length and their tendency to focus on cyberbullying.

We used CNNs that received input text in the form of sequences of integer representations of stemmed unigrams. Our character processing included the conversion of emoticons into word representations, and the removal of non-Latin characters. We also removed frequently occurring url components (e.g., names of popular websites), metadata encoded in the main body-text (e.g., 'RT: '), and a variety of social media platform-specific features. Hashtags and @-mentions were reduced to binary features. The text was then lower-cased and tokenized using NLTK's Tweet-Tokenizer[3]. The tokenized text was next encoded using a dictionary of integers, with the original ordering of the tokens preserved. The encoded text was converted into dense vectors of fixed size. This one-dimensional embedding was fed into a single-layer CNN with 200 embedding dimensions, 150 output dimensions, and 200 convolution kernels. The kernels were optimized using Tensorflow's 'adagrad' optimizer (lr=0.001) using categorical cross-entropy as the loss function. The 150 output dimensions were flattened using a sigmoid function into two output nodes whose values are floats between 0 and 1, with 1 representing *bullying* and 0 representing non-bullying.

To test the performance of our model, we took 70% of the dataset as training set, and 30% of it for testing. As suggested by previous research, we also added textual features (total used: 93) from LIWC 2015 [4] to build another model for comparison. We set the threshold which got the best result (here we used highest F-measure to represent the performance). Comparison can be seen in Table 1. We put ZeroR and SVM (Support Vector Machine) models as baselines for comparison. Because the ZeroR model puts everythoing in the majority class, labeling all of the positive instances as negative ones, both the F-measure for the positive class and True

[3]http://www.nltk.org/api/nltk.tokenize.html

[4]https://liwc.wpengine.com/

Positive score are 0. Meanwhile, the AUC valus of the ZeroR model is 0.5 and the accuracy measure depends on the distribution of positives and negatives in the dataset. It is obvious that the performance of CNN model is better than that of the SVM model in terms of F-measure, AUC, and True Positive rate. It is also expected that adding LIWC features could help to improve the F-measure and accuracy. However, for other important parameters (i.e., True Positives and AUC value), our original model with NLTK-tokenized features got a better index. Note that all the thresholds are taken as 'optimal' because they lead to the highest F-measure, which is not only influenced by True Positive rate but also by the True Negative rate. However, in real-life, we care more about the true positives than the true negatives; in other words, detecting the normal cases (which is much more frequent than cyberbullying) is not the goal for this project. Hence, we are setting up the thresholds with another method which will be described in section 2.3.

Model name	**F-measure**	AUC	**Accuracy**	**True Positive**
ZeroR	0	0.5	87.6%	0%
SVM	0.517	0.851	87.1%	58.9%
SVM + LIWC	0.585	0.892	90.2%	55.9%
CNN	0.523	0.860	87.1%	60.4%
CNN + LIWC	0.597	0.898	90.0%	60.2%

Table 1: Comparison of the results of the CNN models

2.3 Thresholds Setting

To set up the thresholds of our application, we built an electronic survey which contains 45 cyberbullying posts (being labeled as cyberbullying) from the VISR dataset and there are five posts which did not belong to cyberbullying at all (e.g., non-bullying–'he was a complete ass hole, .. He used 2 tell me that my mom tried to abort me because she didn't want to have another kid', cyberbullying–'Go fuck yourself!'). Then we invited four colleagues (two males and two females, all PhD students) to give feedbacks of their feelings about those cyberbullying-related online posts. The survey goes as follow:

"Assume someone (online, you might know the person or not) is sending you a message via phone/posting a message which @yourid/leaving a comment under your profile, etc. Please give the feedback score about how you feel:
1. It is totally fine;
2. Well, not that comfortable, but there is no need to hide it;
3. Not acceptable, I don't want to see this ever, it should be blocked."

After reviewing the feedback scores from our colleagues to ensure they understand the assumption properly, one survey was deleted as the participant didn't get the whole image and returned the feedback with 47/50 'score 3' ("I suggest to hide every F-word, that's really annoying" the participant wrote in the feedback survey). Thus, three surveys were kept for the threshold setting.

The results from the three participants' feelings and the related bullying index of the chosen text are shown in Figure 2. We averaged the 'feeling score' of the three participants; for instance, the average 'feeling score' 1.67 represents the total score of 5, which means that probably two chose 2 and one chose 1. The cyberbullying index of the selected texts is from 0.0431 to 0.8614. We separated the text with 'feeling score' 1 - 1.33 as 'totally fine' group, 1.67 - 2.33 as 'uncomfortable but not that bad' group, 2.34 - 3 as 'not acceptable' group. With an ANOVA test, we found the 'totally fine' group (mean = 0.17) is significantly different ($p = 0.0101$) from the other two groups ('uncomfortable but acceptable': 0.40, 'unacceptable': 0.44).

From the results, six messages are mainly considered as 'totally fine' ('feeling score' from 1 to 1.33, similar to the number of not bullying at all) and the average index is 0.17; we set this index as the threshold_1. From our testing document, most of the 'uncomfortable' and 'not acceptable' scores are higher than threshold_1; only three of them would be identified as 'totally fine'. For the threshold_2, we set it with the average index of 'uncomfortable but not that bad' group (0.40); seven of the 'not acceptable' messages get lower index (which would not be sent with warning, as discuss in the following section), but because of the threshold_1, only one of these 'not acceptable' would not be filtered as 'uncomfortable'.

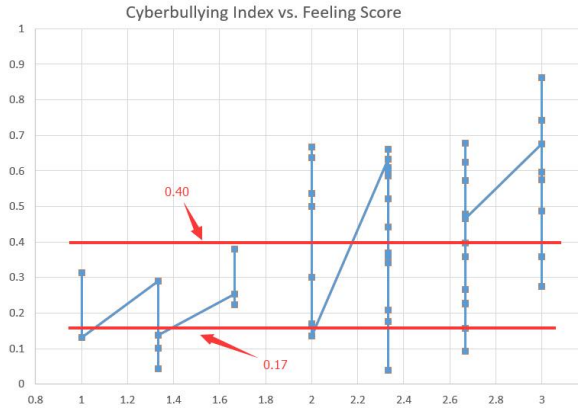

Figure 2: Threshold Setting

3 Application

3.1 Platform and Development Environment

We chose Android OS as our application development platform as it is the most popular mobile OS, and it is an open source, real time operating system which meets our development requirements. We used Android Studio for the development, since it is optimized for all devices, and it provides various APIs and layouts.

3.2 Application Design

3.2.1 Application Subsystem Design

The user interface has a text input field which allows the user to type the message and send it by using the HTTPS to Visr API over the network. This app will extract the cyberbullying index calculated by the Visr API and compare it with the set thresholds. Based on the results of the comparison, a corresponding prompts will be given to the user.

We used HTTPS to transmit data, and we used Volley module to implement the HTTP functions. Volley [5] is a HTTP library developed by Google, it provided us the: 1. Scheduling network request; 2. JSON and images asynchronous downloading; 3. Network request priority handling; 4. Caching; and 5. Powerful APIs.

3.2.2 System Architecture

The system dialog is shown in Figure 3. Regarding users' communication behavior and freedom of speech, the challenges to this interface are presented as follow. On one hand, to build a better online environment, we do not want to miss any of the 'uncomfortable'/'unacceptable' texts and allow them to be sent without filtering. On the other hand, we do not want the users to feel annoying about the intervention; if they really want to send the message, it is not possible to stop them. Hence, we built a win-win solution with the two thresholds and the related interventions.

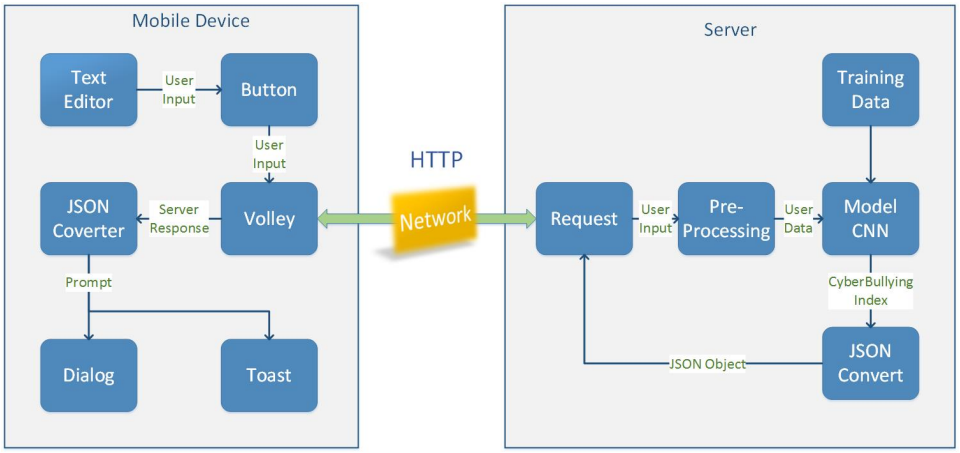

Figure 3: System Diagram

In our interface application, for a text which has index between two thresholds (0.17, 0.40), we send a prompt to the user ('Your message may be aggressive to others, do you really want to send?'). As in this level, the text is not identified as unacceptable; the user could either send it or change it. Similar to figure 1, the prompt is displayed as a delay/chance for user to think about the feelings of the receiver. For the text which has a extremely high index (>0.40) of cyberbullying, the app will first send a prompt as 'Your message may make others feel uncomfortable, please change the tone.' and stay on the same page for the sender to modify the message. If the user

[5] https://developer.android.com/training/volley/index.html

changes it (or not) and the index is still high (above 0.40), the app will send another prompt as 'Your message may make others feel uncomfortable, do you really want to send with warnings?'. This second step comes as we do not want to infringe on users' freedom of speech and we respect the users' communication behavior as well. The only idea is to give advice to the writer about how other people would feel when reading the message. The system flow chart is shown in Figure 4.

4 Examples of the interface

To understand the system better, here are examples of situations with different outputs from the Visr API.

Imagine a user typed a message which receives the cyberbullying index as 0.20 which belongs to the 'uncomfortable but acceptable' level; the prompt would pop us as 'Alert: Your message may be aggressive to others, do you really want to send?' If the user clicks 'yes', it will send the message immediately; if user clicks 'no', there would be a chance to change the message.

For the 'extremely high index' message, such as 'what's your opinion for this fucking shit? You are retard!', the system will display the prompt as shown in figure 5. Please note there is no way to submit this message at the first time. No matter whether the user changes the content or not, if the second time the submitted text's index is still higher than threshold_2, a prompt which is similar to 'uncomfortable but acceptable' would pop up. This prompt is shown in figure 6. If the user wants to send it anyways, there will be warnings with this message, interventions such as hiding the message or sending the message with a warning could be applied at the receiver's end.

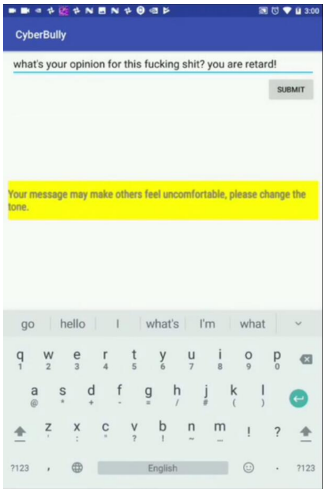

Figure 5: 'Extremely high index' message

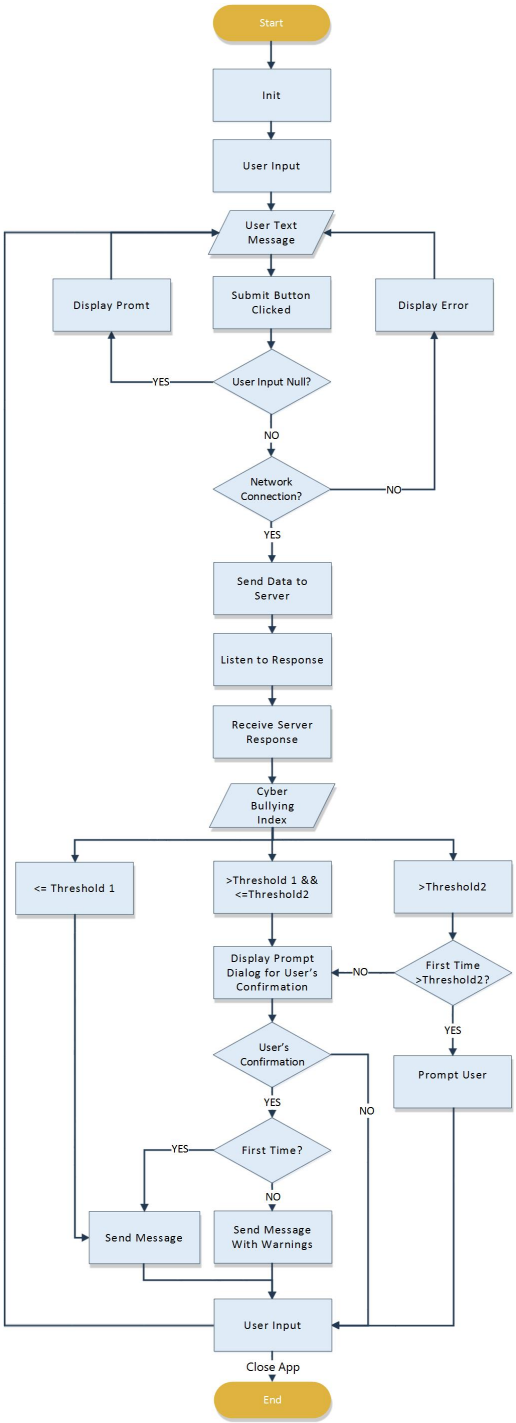

Figure 4: Flow Chart

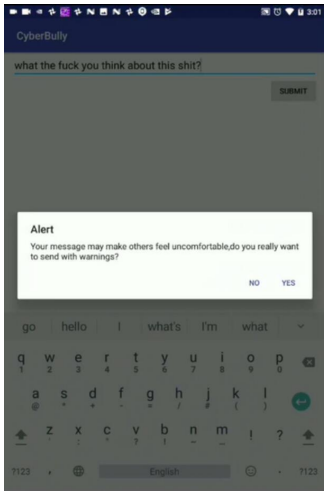

Figure 6: The second time 'extremely high' index of the message

5 Conclusion and Future Work

This paper described the process of creating a cyberbullying intervention application based on a convolutional neural network learning model trained on a multi-platform dataset. The model is then used to compute a cyberbullying index for any new message.

For setting the two thresholds based on the feedbacks of the participants, different interventions could be taken for different levels of the cyberbullying index. Finally, as discussed, this project could be seen as the first step towards a framework of building an effective cyberbullying intervention application for online applications. Social media platforms could use the textual cyberbullying index and our proposed thresholds in order to make interventions to safeguard users from cyberbullying.

Several possible optimizations for future work are as follow:

- Word embeddings, such as GloVe[6] or Word2Vec[7] could be utilized to initialize our CNN models, which might lead to better results.
- As sending images and videos is becoming popular among adolescents (Singh et al., 2017), image/video processing would be another important area for cyberbullying detection.
- The user interface designed in this project takes the user's input directly. In the future, it can be designed as a background running application, which can collect user's input from different applications (while respecting privacy and security issues).
- A complete social network relationship graph (Huang et al., 2014) (for example, whether this conversation is between two good friends or not) could be taken into consideration for improving the cyberbullying identification.

[6]https://nlp.stanford.edu/projects/glove/
[7]https://code.google.com/archive/p/word2vec/

References

Linda Beckman, Curt Hagquist, and Lisa Hellström. 2012. Does the association with psychosomatic health problems differ between cyberbullying and traditional bullying? *Emotional and behavioural difficulties*, 17(3-4):421–434.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 3–6. ACM.

Robin M Kowalski and Susan P Limber. 2007. Electronic bullying among middle school students. *Journal of adolescent health*, 41(6):S22–S30.

Vivek K Singh, Marie L Radford, Qianjia Huang, and Susan Furrer. 2017. " they basically like destroyed the school one day": On newer app features and cyberbullying in schools. In *CSCW*, pages 1210–1216.

Andre Sourander, Anat Brunstein Klomek, Maria Ikonen, Jarna Lindroos, Terhi Luntamo, Merja Koskelainen, Terja Ristkari, and Hans Helenius. 2010. Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of general psychiatry*, 67(7):720–728.

Robert S Tokunaga. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior*, 26(3):277–287.