Automatic Detection of Cyberbullying on Social Networks based on Bullying Features

Rui Zhao
rzhao001@e.ntu.edu.sg

Anna Zhou
annazhou89@gmail.com

Kezhi Mao
ekzmao@ntu.edu.sg

School of Electrical and Electronic Engineering, Nanyang Technological University
Singapore 639798

ABSTRACT

With the increasing use of social media, cyberbullying behaviour has received more and more attention. Cyberbullying may cause many serious and negative impacts on a person's life and even lead to teen suicide. To reduce and stop cyberbullying, one effective solution is to automatically detect bullying content based on appropriate machine learning and natural language processing techniques. However, many existing approaches in the literature are just normal text classification models without considering bullying characteristics. In this paper, we propose a representation learning framework specific to cyberbullying detection. Based on word embeddings, we expand a list of pre-defined insulting words and assign different weights to obtain bullying features, which are then concatenated with Bag-of-Words and latent semantic features to form the final representation before feeding them into a linear SVM classifier. Experimental study on a *twitter* dataset is conducted, and our method is compared with several baseline text representation learning models and cyberbullying detection methods. The superior performance achieved by our method has been observed in this study.

CCS Concepts

•**Information systems** → **Data mining;**

Keywords

Cyberbullying Detection, Text Mining, Representation Learning, Bag-of-Words, Word Embeddings

1. INTRODUCTION

The Internet technology has made a great impact on the communication and relationship among people. In particular, social media platforms now gain their popularity for people of a wide range of ages. Social media provides users not only a good platform for communication and information sharing, but also an easy access to fresh news. However, these platforms are also places where users experience bullying as victims, bullies or bystanders. One study conducted by national anti-bullying charity Ditch the Label in 2013, has shown that two out of three 13-22 years old who were surveyed have been victims of cyberbullying.[1] As reported in [18], approximately 43% of teens once reported being bullied through social media. Another study also shows that cyberbullying victimization rate ranges from 10% to 40% [9].

Different from physical bullying, cyberbullying is "behind-the-scenes" and "24/7". Even worse, the bullying messages left on the Internet will not vanish over time but continuously bother other users. Therefore, the consequences of cyberbullying are even more far-rearching and severe than those of physical bullying.

To detect cyberbullying content underlying huge volumes of posts on social media, a good solution is to develop machine learning-based automatic cyberbullying detection system to categorize the information and generate reports if any cyberbullying is detected, so that all the sensitive information would be modified or erased at the first time, preventing Internet users from overexposure to undesirable information. Natural Language Processing and Text Mining techniques are commonly used in this solution. The framework of machine learning-based automatic cyberbullying detection includes two parts: Representation Learning for Internet Messages and Classification. Each Internet message is firstly transformed into a fixed-length vector and the classifier can be trained in the training corpus over the learned feature space. Finally, the trained classifier can detect the existence of cyberbullying content in each new Internet Message. Similar to other text categorization tasks, the core and vital step is the first step: numerical representation learning. In this field, many previous approaches utilize Bag-of-Words (BoW) model to represent text. BoW model is a classical model in which document is regarded as a multi-set of the words contained in it and modeled as a vector whose weights indicate the occurrence of words in the document. One major limitation of BoW is that each feature corresponding to a term is assumed to be independent to each other and fail to capture semantic information. Feature extraction is always conducted over the BoW features [19]. Besides BoW model, word embeddings, as one exciting fruit of deep learning in NLP community, are able to capture semantic information behind words. In word embeddings, similar words are close to each other, the cosine similarity between word embed-

[1]http://www.express.co.uk/news/uk/433733/69-of-youth-face-cyberbullying

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICDCN '16, January 04-07, 2016, Singapore, Singapore

© 2016 ACM. ISBN 978-1-4503-4032-8/16/01…$15.00

DOI: http://dx.doi.org/10.1145/2833312.2849567

dings reflect the semantic similarity between two words.

In cyberbullying detection, some bullying messages always contain insulting or curse words, which can be regarded as discriminative features. This semantic information can be utilized to provide a robust text representation. Some previous efforts have been made to use these bullying words. Dinakar et.al utilized Linear Discriminative Analysis to learn label specific features [4], which can be regarded as artificial bullying words. In addition, based on prior knowledge, Nahar et.al collected a list of bullying-like words and proposed a weighted TF-IDF scheme by scaling these corresponding bullying-like features with a factor of two [14]. Different from a simple utilization of bullying terms in these previous approaches, our work provides an elaborated use of bullying terms based on word embeddings.

In this paper, we proposed a new representation learning method for cyberbullying detection named embeddings enhanced Bag-of-Words model (EBoW). In EBoW, we firstly define a list of insulting words based on expert knowledge and linguistic resources, i.e., *insulting seeds*. Then, based on word embeddings, we extend *insulting seeds* to define bullying features. Different weights are assigned to bullying features based on the cosine similarity between word embeddings. Considering *insulting seeds* are expanded through word embeddings, the construction of *insulting seeds* do not require much manual labor, which will be illustrated in the section 3.3. Finally, we concatenate the learned bullying features with Bag-of-Words features and Latent semantic features to form the vector representation.

This paper is organized as follows. In section 2, we firstly introduce some related work. Then, we present our proposed Embedding-enhanced BoW (EBoW) model as a representation learning method for cyberbullying detection in Section 3. In Section 4, experimental results on a real *twitter* corpus are illustrated and analyzed. Finally, concluding remarks are provided in Section 5.

2. RELATED WORK

Since our work adopts word2vec embeddings to learn representation for cyberbullying detection, we briefly review some previous works in cyberbullying detection and word2vec embeddings.

2.1 Cyberbullying Detection

The booming of social network leads to the extensive spread of cyberbullying, which is a quite severe problem for children and teenagers. Traditional studies of cyberbullying stand more on a macroscopic view. Conducted by social scientists and psychologists, those studies focus on the statistics of cyberbullying and how to prevent them in a psychological way [11, 6, 8, 5]. As big social network service providers all offer open APIs for academic research, instead of doing statistical study on limited sampled data, researchers are able to access to much larger corpus by using data crawling, which further drives the development of the computational study of cyberbullying based on machine learning and natural language processing techniques. One introductory work has been presented in [17], in which several NLP models such as BoW, Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are applied to detect bullying signals in social media. Their results have verified the possibility of automatic cyberbullying detection. Dinakar et.al used Linear Discriminative Analysis to learn label specific features

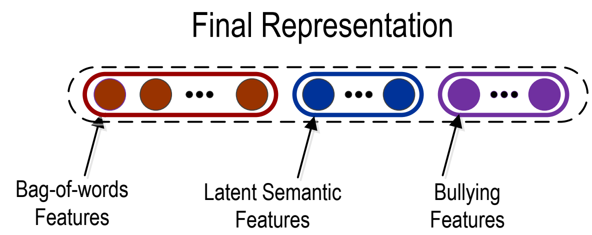

Figure 1: Illustration of final representation in our proposed EBoW model.

and combine them with BoW features to train a classifier [4]. The length of label-specific features is limited to be less than the class numbers, which hinders the performance boost. Nahar et.al magnified the weights corresponding to bullying words by two times [14]. This work shares a similar motivation with the construction of bullying features in our model that bullying features should be enhanced. However, they did not consider the words' semantics and the scaling operation was quite arbitrary. In addition, Nahar et.al [15] also adopted topic models including Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) to learn topics and performed feature selection, which is conducted over topics that feature under bullying-like topics are preserved. However, the determination of bullying-like topics lacks a general theoretical basis.

2.2 Word2vec Embeddings

Recently, Google introduced word2vec, which is an effective and efficient word-embeddings tool [12, 13]. Based on a two-layer neural network language model, word2vec learns vector representations for each word. The tool actually includes two separated models: Continuous Bag of Words (CBoW) and Skip-gram. The training goals of these two methods are reverse. CBoW tries to predict a word given the surrounding words, while Skip-gram tries to predict a window of words given a single word. Due to its surprisingly efficient architecture and unsupervised training protocol, word2vec can be trained over a large-scale unannotated corpus with limited computational resources. The meaningful linguistic relationships between words can be encoded into learned word2vec embeddings.

3. EMBEDDINGS-ENHANCED BAG-OF-WORDS MODEL

In this section, we present our proposed representation learning method for cyberbullying detection. The final representation consists of three parts including Bag-of-Words features, latent semantic features and bullying features based on word embeddings, as shown in Figure 1. These three kinds of features are presented as follows, in particular, the details of bullying features generation.

3.1 Bag-of-Words Features

To extract Bag-of-Words features, a vocabulary including unigram and bigram is constructed firstly and the terms whose document frequencies are less than 2 are all ignored. Different term weighting schemes including *tf-idf* and binary ones can be applied here [3, 16]. In this paper, we adopt the *tf-idf* weighting scheme. The *tf-idf* weight corresponding to the i-th word in j-th document is calculated as follows:

$$w_{i,j} = TF_{i,j} \times log(\frac{N}{DF_i}) \tag{1}$$

where $TF_{i,j}$ is the term frequency of the i-th word in j-th document, DF_i is the number of documents containing i-th word and N is the number of documents.

3.2 Latent Semantic Features

Here, latent semantic features refer to the features extracted by Latent Semantic Analysis (LSA) [10]. In principle, LSA applies Singular Value Decomposition (SVD) on the term-document matrix that each column is the above Bag-of-Words features. Then, the derived latent space is spanned by dominant eigenvectors corresponding to large eigenvalues. Each new feature is a linear combination of all original features.

3.3 Bullying Features

Insulting words can be pre-defined and extracted based on our prior knowledge and other public linguistic resources, which are named *insulting seeds* in this paper. Then, we extend these insulting words automatically based on word embeddings. The insulting words and their corresponding extended terms serve as bullying features.

3.3.1 Insulting Seeds

Since some cyberbullying messages usually contain curse or insulting words, these words are good indications of the existence of bullying. Therefore, we select a list of insulting words based on our prior knowledge and some external linguistic resources[2]. This list contains 350 words indicating curse or negative emotions, such as *nigga, bitch, fuck, slut, whore, twat* etc. Then, we compare the word list with the unigram and bigram features of the corpus used in experiments to obtain the intersection, which is regarded as *insulting seeds* in this paper. For example, the word cloud visualization of *insulting seeds* of the *twitter* dataset used in experimental study is shown in Figure 2. As discussed in the previous work [14], these features are discriminative for cyberbullying detection.

These insulting words are used to construct a vocabulary, and each Internet message is mapped to a vector by counting the occurrence times of each word appears. Different term weighting schemes including *tf-idf* and binary ones are applied. However, such a direct use of these insulting words could hardly achieve robust representation. The reasons are two-fold. Firstly, due to the linguistic complexity and the synonymy, the obtained insulting words can not cover all the possible bullying terms. Secondly, effect intensities of different insulting words are different. The occurrence of *slut* is more indicative of the existence of cyberbullying than that of *nerd*, although these two words both express negative affection. To overcome these two limitations, we employ word embeddings to extend these insulting words and weight them.

3.3.2 Bullying Features Generation based on Word Embeddings

[2] A collection of insulting words can be found in the website: http://www.noswearing.com/dictionary

Figure 2: Word cloud visualization of our pre-defined insulting words, i.e., *insulting seeds*.

Different from the one-hot representation in Bag-of-Words model, word embeddings use real-valued and low-dimensional vectors to represent semantics of words. In the vector space behind well-trained word embeddings, similar words are placed close to each other. For example, the words *China* and *Singapore* are close to each other, and the words *beef* and *pork* are close to each other. The cosine similarity between word embeddings can represent the semantic similarity between the two corresponding words. Since we focus on *twitter* corpus here, we utilize a well-trained word2vec model on a large-scale twitter corpus including 400 million tweets [7]. A visualization of some word embeddings based on PCA is shown in Figure 3. It is observed that curse words form distinct clusters, which are also far away from normal words. Even insulting words are located at different regions due to different word usages and insulting expression.

We extend the pre-defined *insulting seeds* based on word embeddings. For each *insulting seed*, we select the top-h most-similar words in the vocabulary as extended bullying features. Based on word embeddings, the cosine similarity of the two embeddings can measure the similarity between these two words. It should be noted that the vocabulary of our corpus consists of unigrams and bigrams. For one bigram $w_l w_r$, we simply use an additive model to derive the corresponding embedding as follows:

$$\mathbf{v}(w_l w_r) = \mathbf{v}(w_l) + \mathbf{v}(w_r) \tag{2}$$

To give a clear illustration, the top-10 similar terms to the *insulting seed*: "*slut*" and their cosine similarities scores are shown in Table 1. These cosine similarity scores can be used to weight these extended features. And the weights of these *insulting seeds* are set to 1. These extended features and *insulting seeds* compose the final bullying features. It should be noted that one word may be similar to several *insulting seeds*. To address the possible overlapping, we only keep the highest cosine similarity of the word as its weight. Therefore, the length of the final bullying features is less than or equal to $n * h$ where n is the number of *insulting seeds*.

3.4 Feature Concatenation

Table 1: **Top-10 similar terms to the insulting word: *slut*. They are retrieved based on word embeddings. Their corresponding cosine-similarity scores are also shown.**

Similar Words	Cosine Similarity Scores
a slut	0.815
whore	0.738
a whore	0.638
hypocrite	0.536
bitch	0.508
puta	0.468
nerd	0.455
bully her	0.451
fat bully	0.440
bully nigga	0.435

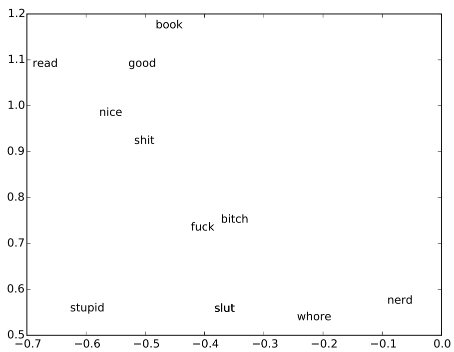

Figure 3: Two dimensional visualization of our used Twitter word embeddings via PCA. Displayed terms include both bullying ones and normal ones. It shows that similar words are nearby vectors.

The three above kinds of features including Bag-of-Words features, latent semantic features and bullying features are concatenated to form the final representation. When final representation for each Internet message is obtained, a classifier is built to detect the existence of cyberbullying. In our proposed system, we feed learned features into a linear SVM pattern classifier.

4. EXPERIMENTS

In this section, we evaluate our proposed EBoW model using a real-world cyberbullying dataset. The details of our adopted dataset and experimental settings are introduced firstly. Experimental results and comparison with other baseline methods are followed.

4.1 Descriptions of Datasets

The public bullying traces dataset is adopted here[3], which consists of tweets that are messages sent on *Twitter* [17]. *Twitter* is a very popular social media application, which is able to let you access to the latest stories, ideas, opinions and news(https://about.twitter.com/). Registered users can read and post tweets, which are defined as the messages posted on *Twitter* with a maximum length of 140-character. Twitter has over 270 million active users, and 500 million tweets are posted per day. Behind these amazing statistics, cyberbullying has sadly become a common occurrence. According to a previous study[4], there were about 100,000 abusive tweets a week. Therefore, data crawled from *Twitter* is a good source for cyberbullying research.

The *Twitter* dataset is composed of tweets crawled by the public *Twitter* stream API. Each tweet contains at least one of the following keywords: bully, bullied, bullying. Retweets are removed by excluding tweets containing the acronym ''RT''. Finally, 1762 tweets are sampled uniformly from the whole tweets collections on August 6, 2011 and manually labeled. It should be pointed out here that labeling is based on bullying traces. Bullying traces are defined as response to the bullying experience, which include but far exceed the incidences of cyberbullying. Some examples of bullying traces are shown in Figure 4. To preprocess these tweets, a twitter-specialized tokenizer[5] is applied without any stemming or stopword removal operations. In addition, some special characters including user mentions, URLS and so on are replaced by predefined characters, respectively. The statistics of this dataset can be found in Table 2. Since the dataset does not have explicit train/test split, 5-fold cross validation (CV) is applied, where four-folds are used for training and the remaining one is used for testing. The mean results will be reported.

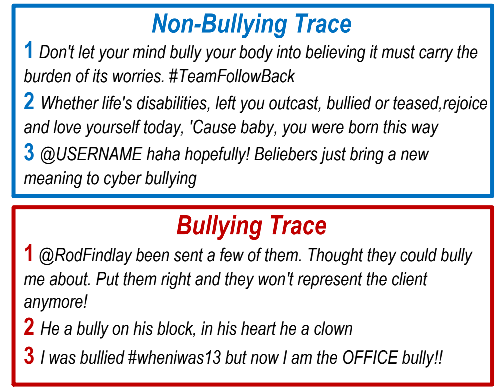

Figure 4: Some examples from *Twitter* Datasets. Three of them are non-bullying traces. And the other three are bullying traces.

Table 2: Statistical properties of the *Twitter* dataset.

Statistics	*Twitter*
Sample No.	1762
Bullying Instances	684
Train/Test	5-fold

4.2 Experimental Setup

[3]The dataset has been kindly provided at http://research.cs.wisc.edu/bullying/data.html

[4]http://news.wisc.edu/20931

[5]http://www.ark.cs.cmu.edu/TweetNLP/

The following methods will be compared:

* BoW Model: the raw BoW features are directly fed into the classifier.

* Semantic-enhanced BoW Model: This approach is referred in [14]. Following the original setting, we scale the bullying features by a factor of 2. To give a fair comparison, the bullying features here is *insulting seeds* used in our proposed method.

* LSA: Latent Semantic Analysis [10].

* LDA: Latent Dirchilet Allocation [1]. Our implementation of LDA is based on *Gensim*[6].

* EBoW: Our proposed Embeddings-enhanced Bag-of-Words Model.

For BoW, the most frequent 2000 terms including unigram and bigram are used as features. For LSA and LDA, the number of latent topics are both set to 100. To implement LDA, we set hyperparameter α for document topic multinomial and hyperparameter η for word topic multinomial to 1 and 0.01, respectively.

For our proposed method, BoW and latent semantic features are both the same as those in two above compared models. For bullying features, the number of *insulting seeds* is 20. The number of similar words compared to each insulting seed h is set to 50. After feature expanding based on word embeddings, the length of bullying features is 641 finally. Therefore, the dimension of features learned by our EBoW is the sum of the lengths of the three kinds of features as: $2000 + 100 + 641 = 2741$.

We then apply linear SVM [2] in the new feature space generated by the above mentioned approaches. In linear SVM, the hyper-parameter: regularization term C, is searched over a range as $\{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3, 2^4, 2^5\}$ via a five-fold cross-validation conducted on training data.

The cyberbullying detection task is thus converted into a binary classification problem. The adopted evaluation metrics include precision, recall and F1-value.

4.3 Experimental Results

We reported the precision, recall and F1 scores of all compared methods in Table 3. Obviously, our proposed method outperforms other compared methods in all the three evaluation metrics.

It is observed that semantic BoW model is able to give a slight better performance than BoW. In sBoW, the weights corresponding to bullying features are multiplied by 2 and the bullying features are the same as the *insulting seeds* in our proposed EBoW. This indicates that bullying features can indeed improve cyberbullying detection. Different from a direct use of *insulting seeds* in sBoW, our approach EBoW expands the pre-defined insulting words through word embeddings and assign different weights to these expanded bullying features according to cosine similarity between word embeddings. As a result, our approach is able to gain a significant performance improvement compared to sBoW over all three evaluation measures.

We also observe that our method outperforms two stat-of-arts text representation learning methods: LSA and LDA.

[6]https://radimrehurek.com/gensim/index.html

Table 3: Precision (%), Recall (%) and F1 Scores (%) for compared methods. Bold face indicates best performance.

Measures	BoW	sBoW	LSA	LDA	EBoW
Precision	75.6	75.7	75.9	74.0	**76.8**
Recall	77.8	78.3	78.2	76.5	**79.4**
F1 Score	76.6	76.9	77.0	74.9	**78.0**

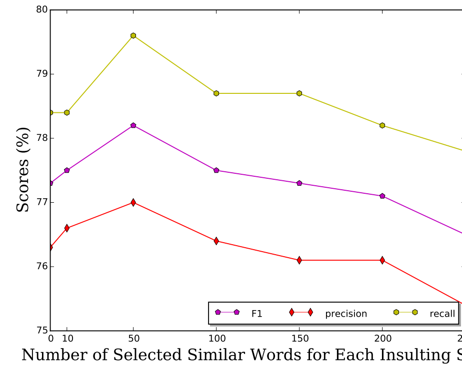

Figure 5: Performance of our proposed EBoW model for different h. Different colors indicate different measures.

This may be because that both methods belong to dimensionality reduction. The reduced dimension is a key parameter to determine the quality of learned feature space. Here, we fix the dimension of latent space to 100. A deliberate searching for this parameter may boost the performances of LSA and LDA

4.4 Parameters Sensitivity

In our proposed EBoW model, we select the top-h similar words for each pre-defined insulting word. Parameter h is used as the threshold to expand bullying features. An increased h will mean more terms are considered as bullying features so that the dimension of bullying features increases. In this section, we investigate the influence of the parameter on our model performance for cyberbullying detection. Parameter h is chosen from a predefined set: [0, 50, 100, 150, 200, 250]. When $h = 0$, bullying features only consist of our pre-defined insulting words. According to the total 6 settings of h, we feed our corresponding learned EBoW features into the classifier. The other experimental settings are kept unchanged. Then, precision, recall and F1 scores for these different settings are calculated and reported in Figure 5.

It shows that a moderate h is able to achieve the best performance. If h is too small, some discriminative terms that are not covered by the pre-defined insulting words are filtered out in the learned bullying features. If h is too large, some irrelevant terms are considered as bullying features. Both scenarios will lead to an ineffective feature space for cyberbullying detection. As a result, a classifier applied on such a feature space may not produce satisfying performance.

5. CONCLUSIONS

In this paper, we have proposed a novel representation learning method for cyberbullying detection, which is named Embedding -enhanced Bag-of-Words. EBoW concatenates BoW features, latent semantic features and bullying features together. Bullying features are derived based on word embeddings, which can capture the semantic information behind words. When the final representation is learned, a linear SVM is adopted to detect bullying messages. The effectiveness of our proposed model has been experimentally verified through a real-world *Twitter* corpus.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[3] G. Chowdhury. *Introduction to modern information retrieval*. Facet publishing, 2010.

[4] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *The Social Mobile Web*, 2011.

[5] M. Fekkes, F. I. Pijpers, A. M. Fredriks, T. Vogels, and S. P. Verloove-Vanhorick. Do bullied children get ill, or do ill children get bullied? a prospective cohort study on the relationship between bullying and health-related symptoms. *Pediatrics*, 117(5):1568–1574, 2006.

[6] G. Gini and T. Pozzoli. Association between bullying and psychosomatic problems: A meta-analysis. *Pediatrics*, 123(3):1059–1065, 2009.

[7] F. Godin, B. Vandersmissen, W. De Neve, and R. Van de Walle. Named entity recognition for twitter microposts using distributed word representations. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 146–153, Beijing, China, July 2015. Association for Computational Linguistics.

[8] J. Juvonen and E. F. Gross. Extending the school grounds?âĂŤbullying experiences in cyberspace. *Journal of School health*, 78(9):496–505, 2008.

[9] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. 2014.

[10] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[11] B. L. McLaughlin, A. A. Braga, C. V. Petrie, M. H. Moore, et al. *Deadly Lessons:: Understanding Lethal School Violence*. National Academies Press, 2002.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[14] V. Nahar, X. Li, and C. Pang. An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 2012.

[15] V. Nahar, X. Li, and C. Pang. An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 3(5):238, 2013.

[16] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)*, 26(3):13, 2008.

[17] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore. 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. Association for Computational Linguistics, 2012.

[18] M. Ybarra. Trends in technology-based sexual and non-sexual aggression over time and linkages to nontechnology aggression. *National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda*, 2010.

[19] R. Zhao and K. Mao. Semi-random projection for dimensionality reduction and extreme learning machine in high-dimensional space. *Computational Intelligence Magazine, IEEE*, 10(3):30–41, 2015.