SOSNet: A Graph Convolutional Network Approach to Fine-Grained Cyberbullying Detection

Jason Wang[1], Kaiqun Fu[2], Chang-Tien Lu[2]
[1]*Thomas Jefferson High School for Science and Technology, Alexandria, VA, USA*
[2]*Virginia Tech, Falls Church, VA, USA*
jasonwang292@gmail.com, {fukaiqun, ctlu}@vt.edu

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9378065

***Abstract*—Amidst the COVID-19 pandemic, cyberbullying has become an even more serious threat. Our work aims to investigate the viability of an automatic multiclass cyberbullying detection model that is able to classify whether a cyberbully is targeting a victim's age, ethnicity, gender, religion, or other quality. Previous literature has not yet explored making fine-grained cyberbullying classifications of such magnitude, and existing cyberbullying datasets suffer from quite severe class imbalances. To combat these challenges, we establish a framework for the automatic generation of balanced data by using a semi-supervised online Dynamic Query Expansion (DQE) process to extract more natural data points of a specific class from Twitter. We also propose a Graph Convolutional Network (GCN) classifier, using a graph constructed from the thresholded cosine similarities between tweet embeddings. With our DQE-augmented dataset, which we have made publicly available, we compare our GCN model using eight different tweet embedding methods and six other classification models over two sizes of datasets. Our results show that our proposed GCN model matches or exceeds the performance of the baseline models, as indicated by McNemar statistical tests.**

***Index Terms*—cyberbullying, dynamic query expansion, graph convolutional network, social media data mining, machine learning**

I. Introduction

As social media usage becomes increasingly prevalent in every age group,[1] a vast majority of citizens rely on this essential medium for day-to-day communication. Social media's ubiquity means that cyberbullying can effectively impact anyone at any time or anywhere, and the relative anonymity of the internet makes such personal attacks more difficult to stop than traditional bullying. The COVID-19 pandemic notably makes cyberbullying an increasingly worrying threat; on April 15th, 2020, UNICEF issued a warning in response to the increased risk of cyberbullying during the COVID-19 pandemic due to widespread school closures, increased screen time, and decreased face-to-face social interaction.[2]

Cyberbullying, as discussed by [1], has been defined as "the use of digital technology to inflict harm repeatedly or to bully," but its digital nature and relative anonymity make it difficult to specifically pinpoint the imbalance of power, repetition, and harmful intent that is typically associated with traditional bullying, and manifests as a still-ongoing research topic. The statistics of cyberbullying are outright alarming: ***36.5%*** of middle and high school students have felt cyberbullied and ***87%*** have observed cyberbullying, with effects ranging from *decreased academic performance* to *depression* to *suicidal thoughts*.[3]

The methods currently in place to combat cyberbullying primarily consist of teaching "Internet street smarts," looking for warning signs, and counseling [2]. Legal consequences for cyberbullying are present in some form in all 50 US states (although not at the federal level), yet the majority of these laws have limited to zero jurisdiction outside of school.[4] Facebook, Twitter, Instagram, Snapchat, and other big name social media platforms have cyberbullying guides/resources and passive reporting mechanisms built-in to their application; however, they have yet to provide active anti-cyberbullying functions. An active system is absolutely critical because an estimated 90% of cyberbullying activities go unreported.[5] Multiple awareness groups exist for cyberbullying, but despite these efforts, cyberbullying attacks are still increasing in number.[6]

In this work, we seek to improve *active machine learning-powered cyberbullying detection* on Twitter, specifically focusing on the ability to discern what specific quality of the victim the cyberbully is attacking: **age, ethnicity, gender, religion, or other**. By actively targeting individual malicious tweets, we hope to stop the inflicted harm before it builds up to a cyberbullying level, and by providing a fine-grained classification of such harmful tweets, we hope to better inform users and counselors for a more targeted healing process. Automatic cyberbullying detection fits into the broader field of sentiment analysis in natural language processing, a major application in machine learning. The textual data is converted into a representative mathematical vector that downstream classifiers can take as input through conversion techniques such as TF-IDF, word2vec, and BERT [3]. The challenge behind sentiment analysis is the ability to detect the subtle nuances and hidden subtexts in language, such as sarcasm, irony, metaphors, allu-

[1]https://www.pewresearch.org/internet/fact-sheet/social-media/

[2]https://www.unicef.org/press-releases/children-increased-risk-harm-online-during-global-covid-19-pandemic

[3]https://www.broadbandsearch.net/blog/cyber-bullying-statistics

[4]https://cyberbullying.org/bullying-laws

[5]https://www.webmd.com/parenting/features/prevent-cyberbullying-and-school-bullying

[6]https://cyberbullying.org/summary-of-our-cyberbullying-research

sions, idioms, nicknames, double negatives, and word order. Vocabulary alone may not be sufficient in detecting cyberbullying; automatic cyberbullying detection will depend heavily on the quality of current natural language understanding models to generate representative word and sentence embeddings. An additional challenge is that while cyberbullying behavior is negative, the concept of cyberbullying does not fit cleanly into the typical positive/neutral/negative polarity task of sentiment analysis. Our research focuses on answering three overarching research questions: 1) Is Dynamic Query Expansion (DQE) a viable alternative to oversampling/downsampling? 2) Given that prior authors have shown success in differentiating cyberbullying and not cyberbullying, can we show similar progress for the problem of fine-grained cyberbullying classification, where all tweets share a broad hateful sentiment? 3) How can a graph input for a Graph Convolutional Network (GCN) be constructed for cyberbullying detection, and does a graph structure provide more insight to the problem than traditional machine learning classifier models?

Our motivations for studying DQE stem from the challenges facing current cyberbullying research, namely, the lack of balanced datasets and the lack of publicly available finely labeled cyberbullying data, as prior research has mainly focused on binary or ternary classification. We have also noticed severe class biases in previous cyberbullying datasets (see Section IV). Another challenge is the sparsity of textual data; we would like to utilize a graph structure that will provide new insights to how ideas are connected, and therefore aid this tougher fine-grained analysis. If we can construct a neighbors graph within the realm of cyberbullying out of sentence embeddings (with the aim of encoding a primitive "knowledge" or "conceptual" graph), a GCN will be able to take advantage of those connections. In this paper, we propose a **S**emantic C**o**sine **S**imilarity Graph Convolutional **Net**work (SOSNet) to address these challenges. Our main contributions of this work are:

- **Developing an online Dynamic Query Expansion process using concatenated keyword search.** We improve upon the DQE algorithm by establishing a procedure to iteratively expand an existing dataset by connecting the algorithm's output with Twitter through GetOldTweets3. We successfully leverage the algorithm to generate new data points of a narrow class in a semi-supervised fashion. This procedure automates most of the data collection process as well as facilitates the collection of balanced data, solving the class imbalance problems faced by almost every other dataset.
- **Formulating a graph structure of tweet embeddings and implementing a Graph Convolutional Network for fine-grained cyberbullying classification.** GCNs are a previously unexplored classification method in the field of cyberbullying detection. We propose a GCN-based framework that takes in a graph generated by the thresholded semantic cosine similarities between every tweet, allowing for the effective propagation of labels across tweets with similar main ideas. This helps combat the sparsity of textual data, and we show promising results for its application in cyberbullying classification.
- **Curating a balanced multiclass cyberbullying dataset from DQE, and making it publicly available.**[7] Our multiclass dataset enables future research into fine-grained cyberbullying classification, a previously unexplored area. Researchers interested in binary cyberbullying classification will also find our dataset useful as the balanced multiclass labels offer more representative samples of cyberbullying, allowing for better generalization of cyberbullying over the spectrum of offensive messages, in contrast with current datasets containing cyberbullying texts of a limited scope.
- **Evaluating a combination of eight tweet embedding methods and seven classification models on two sizes of data for this fine-grained cyberbullying classification task.** Differentiating between five types of cyberbullying is a novel task, so we performed an exhaustive search over the many existing methods to highlight which of them translate over to this task well, and set the baselines for future research. By meticulously experimenting with so many comparison methods, we showcase the performance of our SOSNet in full context.

The rest of our paper is structured in the following manner: we review related works in Section II. In Section III, we express our problem setup, and in Section IV, we present our proposed modified Dynamic Query Expansion algorithm and Graph Convolutional Network framework. Section V contains our extensive experiments and thorough analyses of our results. We conclude our paper with a summary and discussion of what our research represents in this field and to our community.

II. Related Works

In this section, we provide a review of current research on social media-based cyberbullying detection. We break this topic into three subtopics: cyberbullying detection via social media analysis, spatiotemporal event detection on Twitter, and Dynamic Query Expansion for event detection.

Cyberbullying detection via Social Media Analysis. The study of cyberbullying detection has gained increasing attention in recent years due to its harmful influences on society. A growing body of research is emerging in advanced techniques to detect cyberbullying-related activities. Most recent works utilize advanced machine learning and natural language processing approaches to identify the cyberbullying exchanges by searching for textual patterns representative of the verbally abusive activities online. Dinakar et al. [4] performed experiments with classifiers such as Naïve Bayes and SVM on a set of messages clustered by themes and found performance to be much improved on individual clusters over the combined set. Dadvar and De Jong [5] and Dadvar et

[7]Dataset (Warning: Explicit Content) can be found at https://drive.google.com/drive/folders/1oB2fan6GVGG83Eog66Ad4wK2ZoOjwu3F?usp=sharing

al. [6] adopted an approach of training an SVM classifier on MySpace posts grouped by the users' gender. Their discovery shows that cyberbullying detection was significantly improved on the gender-grouped posts when compared against results obtained when the same classifier was trained on non-grouped datasets. Chavan and Shylaja [7] proposed an algorithm to calculate a score representing the probability of a comment being offensive to other users. They utilized a selection of features, such as skip-grams and combining the results of SVM and Logistic Regression classifiers. Squicciarini et al. [8] applied decision tree classifiers on content-specific features to detect cyberbullies in social networks such as MySpace and spring.me, and additionally proposed a rule-based algorithm to further detect cyberbullying behaviors.

Spatiotemporal Event Detection on Twitter. Our work is also generally related to social media event detection [9], [10] using Twitter. This field covers various events such as natural disasters [11], criminal incidents [12], disease outbreaks [13], population migrations [14], trending news [15], [16], and activity planning [17]. One common method for event extraction is to use unsupervised learning models that work via keyword matching, clustering, and topic modeling [18]–[20]. Some example applications include detecting incidents of civil unrest [21] and imminent terrorist threats to airports [22]. Researchers have also used supervised learning models on social media data for stock market predictions [23], crime predictions [12], and civil unrest detection [21].

Dynamic Query Expansion for Event Detection. Query expansion is a process that reformulates the seed query in order to improve the coverage and accuracy of information retrieval [24]. To improve the performance of this retrieval in Twitter, a new thread of work utilizes query expansion to dynamically expand keywords [25], retrieve tweets [26], and discover events [27]. Previously, Khandpur et al. [22] used DQE to find terrorist threats against airports, Zhao et al. [28] used DQE to monitor flu outbreaks, and Zhao et al. [27] used DQE to collect specific details about civil unrest in Latin America. In the domain of cyberbullying detection, Chatzakou et al. [29] used an "automatic snowball sampling" that is almost identical to the DQE algorithm; they use a dynamic list, a ranking process, and updates over multiple time steps. Their usage, however, was different from both the other DQE studies and our study because they (i) focused on collecting a dataset from scratch, (ii) considered only hashtags, and (iii) did not mention the use of TF-IDF for ranking, only frequency. Other cyberbullying studies [30]–[35] use a keyword search using a list of expletives, forms of bullying, or mentions of school. This approach will simply not work for fine-grained classification, as the resulting searches would be extremely noisy with general hateful and not hateful tweets that would be manually laborious to sort. This is the main reason we opted to explore a semi-supervised version of DQE.

III. Problem Statement

In this section, we first provide the mathematical definition of the cyberbullying detection problem. Next, we describe the building blocks for the topic detection with our proposed Dynamic Query Expansion (DQE) process. They collaborate to construct our proposed SOSNet model.

A. Cyberbullying Detection with GCN

Assume that we are given a set of online posts $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) \in \mathbb{R}^{N \times F}$ from the social media platforms, where N is the total number of posts in our input data and F is the number of features that preserve the semantic meanings of the posts.

Definition I: *Textual graph*. We consider each online post $\mathbf{x}_i$ as one node in the graph and construct a fully-connected graph representation $\mathbf{G}^* = (\mathbf{V}, \mathbf{E}^*)$ of the given online posts dataset $\mathbf{X}$, where $|\mathbf{V}| = N$, and $\mathbf{V}$ is the corresponding vertex set for the online posts set $\mathbf{X}$; $\mathbf{E}^*$ is the edge set for the fully connected graph $\mathbf{G}^*(|\mathbf{E}^*| = \binom{N}{2})$. When the filtering criteria ϵ holds, a textual graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ representing the semantic distance between the posts is constructed, where $\mathbf{E} \subseteq \mathbf{E}^*$.

We also define a vector $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N) \in \{b_k | k = 1, 2, ..., K\}^N$, where b_k is the k^{th} cyberbullying class in the label dataset; K is the number of targeted cyberbullying classes. With such concepts introduced, our cyberbullying detection problem is defined as follows: given the input data $\mathbf{X}$, the filtering criteria ϵ, and the corresponding labels Y, can we find an optimal solution to accurately infer the type of cyberbullying activities when given an unseen verbally abusive online post? Mathematically, the problem can be formulated as learning a function $\mathbf{F}^*$ which maps $\mathbf{X}$ to $\mathbf{Y}$:

$$\mathbf{F}^*(\mathbf{X}) \rightarrow \mathbf{Y} \tag{1}$$

The problem is challenging in three aspects: 1) Features F and the data samples are within the same order of magnitude, which implies that this is a high-dimensional setting and therefore likely to exhibit sparsity. Indeed, the sparsity is likely more severe due to the social media platforms' restrictive character count. 2) Certain cyberbullying activities within some groups of users (e.g. religion, age) are not as widely discussed on social media platforms. These weak signals are hard to capture. 3) The relatedness across different types of cyberbullying activities varies in feature space and is too crucial to be neglected.

B. Dynamic Query Expansion for Cyberbullying Detection

In order to overcome the challenge of biased numbers of cyberbullying activities, we adopt an online query expansion method to further augment the initial cyberbullying dataset. The input to our Dynamic Query Expansion is a collection of initial online posts $\mathbf{X}^k = (\mathbf{x}_1^k, \mathbf{x}_2^k, ..., \mathbf{x}_{p_k}^k)$ where p_k is the initial number of online posts for the k^{th} target cyberbullying class b_k. We define the initial size of the dataset as $P = \sum_{k=1}^{K} p_k$. As we defined in the previous subsection, $\mathbf{X} \in \mathbb{R}^{N \times F}$ denotes the dataset we require, thus the number of online posts expanded by our Dynamic Query Expansion is $N - P$. Let $\mathbf{X}_+^k$ denote the subspace of the target online posts (in our case, the posts containing the relevant cyberbullying queries).

Definition II: *Seed Query*. A seed query $\mathbf{Q}_0$ is a manually selected and typed dependency query targeted for a certain type of event. For instance, "ni**er" can be defined as a potential seed query for a cyberbullying activity targeting certain ethnic groups.

Definition III: *Expanded Query*. An expanded query $\mathbf{Q}_k$ is a typed dependency query that is automatically generated by the Dynamic Query Expansion algorithm based on a set of seed queries and a given online posts collection $\mathbf{X}^k$. The expanded query and its seed query can be two different descriptions of the same subject. More commonly, an expanded query can be more specific than its seed query.

DQE Task: Given a small set of seed queries $\mathbf{Q}_0$ and an initial online post collection $\mathbf{X}^k$, the task of online Dynamic Query Expansion is to iteratively expand $\mathbf{X}^k_+$ and $\mathbf{Q}_k$ until all the target-related online posts are included.

IV. Methodology

In this section, we present our proposed model, SOSNet, which tackles the problem of predicting the type of cyberbullying activities. First, we discuss the construction process of the textual graph for our proposed SOSNet model. We also show an overview of the design of the proposed framework. Then, we detail the architecture of the Graph Convolution Networks used on the two major building blocks of SOSNet: *Textual Graph Construction*, and the *Online Dynamic Query Expansion* of our model. These two building blocks work consecutively on extracting online posts of the target domains and learning the inference function of the cyberbullying detection. Then, we describe the training procedure, which unifies the distinct components of our proposed framework.

A. Textual Graph Construction

The textual graph structure captures the semantic correlations and similarities between the online posts on the target social media platforms, which provide local textual perception. The textual graph is constructed based on the text similarities between the online posts, which provide a global perspective.

According to Definition I, we describe the textual graph with a partial graph representation $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, $\mathbf{A}$ denotes the adjacency matrix of the graph, and $\mathcal{V}$ is the set of vertices in the dual graph which represents the set of online posts $\mathbf{X}$. For online post $\mathbf{x}_i$ and $\mathbf{x}_j$ which are represented by vertices v_i and v_j respectively in $\mathcal{G}$, the adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ of the graph can be formulated as:

$$\mathbf{A}_{i,j} = \begin{cases} \langle \mathbf{x}_i, \mathbf{x}_j \rangle, & \text{if condition } \zeta(\mathbf{x}_i, \mathbf{x}_j) \text{ holds.} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ is the cosine similarity calculation between the vectors; ζ is a textual similarity condition that calculates the textual similarity between the input targets $\mathbf{x}_i$ and $\mathbf{x}_j$. In the SOSNet model, we select the ζ condition as: $\langle \mathbf{x}_i, \mathbf{x}_j \rangle \geq \epsilon$. Such textual condition only returns true if $\mathbf{x}_i$ and $\mathbf{x}_j$ have a meaningful semantic similarity to each other, as ϵ is empirically chosen to trim as many insignificant edges as possible and reduce the complexity of the network while preserving the performance of the model.

B. Convolution on Textual Graph

Given the textual graph representation of the online posts $\mathbf{X}$ and $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, the spectral graph convolution is operated in the Fourier domain. An essential operator for spectral graph analysis is the Laplacian matrix $\mathbf{L}$, which is defined by $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D} \in \mathbb{R}^{N \times N}$ is the degree matrix, and $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{r_{ij}}$. With the definitions of the degree matrix and the Laplacian matrix, we further calculate the normalized Laplacian matrix by $\mathbf{L}_n = I_N - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \in \mathbb{R}^{N \times N}$, where I_N is the identity matrix. After the normalization process, the Laplacian matrix $\mathbf{L}_n$ is now symmetric positive semidefinite; its spectral decomposition is represented as $\mathbf{L}_n = \mathbf{U}\Lambda\mathbf{U}^T$. $\mathbf{U}$ is comprised of orthogonal and normalized eigenvectors $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_N] \in \mathbb{R}^{N \times N}$ and $\Lambda = diag([\lambda_1, ..., \lambda_N])$ is the combination of eigenvalues $\lambda \in \mathbb{R}^N$. Then, the spectral convolution can be defined in the Fourier domain as:

$$y = \sigma(\mathbf{U} g_\theta(\Lambda)\mathbf{U}^T x) \tag{3}$$

where x, y are the convolution input and output respectively, g_θ is the filter of the convolution process, and σ is the activation function. This formation is feasible for the spectral convolution process; however, it requires expensive computation complexity for large scale graph structures. To reduce the computation complexity, an approximation should be applied to the filter $g_\theta(\Theta)$. We apply m^{th} order Chebyshev polynomials $T_m(x)$ to the filter $g_\theta(\Theta)$:

$$g_\theta(\Lambda) \approx \sum_{m=0}^{K} \theta_m T_m(\widetilde{\Lambda}) \tag{4}$$

$$\widetilde{\Lambda} = \frac{2}{max(\lambda)}\Lambda - I_N \tag{5}$$

This approximation was first proposed by Hammond et al. [36]. The Chebyshev polynomials are recursively defined as $T_m(x) = 2xT_{m-1}(x) - T_{m-2}(x)$, with $T_0(x) = 1$ and $T_1(x) = x$. Kipf et al. [37] further limit the number of order m to be 1, along with the max eigenvalue to be 2. The Graph Convolution Network is now represented as:

$$\mathbf{Y} = (\mathbf{D} + I_N)^{-\frac{1}{2}}(\mathbf{A} + I)(\mathbf{D} + I_N)^{-\frac{1}{2}} X\Theta \tag{6}$$

For the prediction stage, the target is the similarly formulated textual graph $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ where $\mathcal{V}^*$ additionally contains the new online post observations up for prediction and $\mathcal{E}^*$ is the thresholded semantic similarities between the vertices of the updated $\mathcal{V}^*$. The input $X \in \mathbb{R}^{N \times K}$ is the hidden textual features generated by the sentence-based post encoders, and the output $Y \in \{b_k | k = 1, 2, ..., K\}^N$ is the predicted cyberbullying labels based on the connectivity of the online posts. Information sharing between the connected nodes can be modeled by the filter g_θ. Thus SOSNet can be utilized as an appropriate model for classifying the cyberbullying activities.

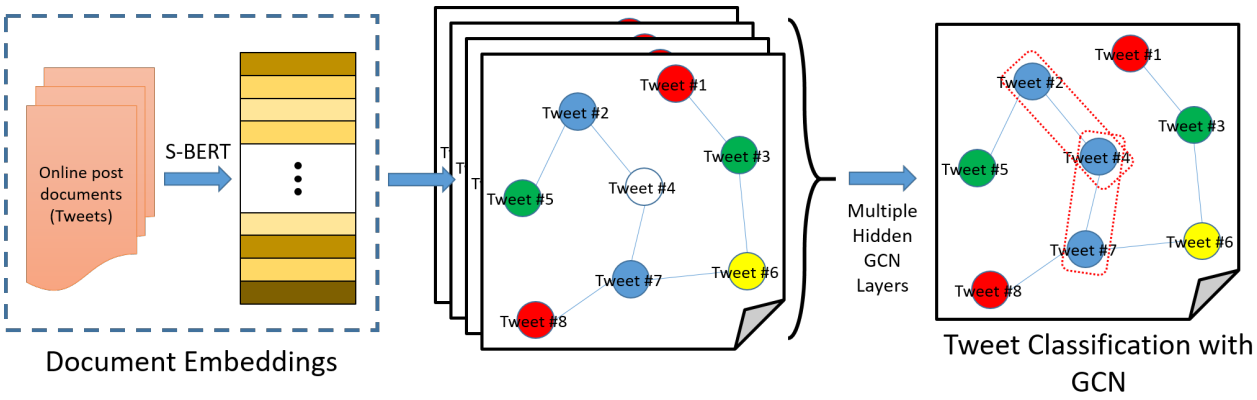


Fig. 1: Our Proposed SOSNet Framework for Fine-Grained Cyberbullying Classification

C. Online Dynamic Query Expansion

Dynamic Query Expansion (DQE) is commonly used as a data mining technique. Data mining is a process that extracts usable patterns and other information from a large pool of data. It is widely used to make sense of the hundreds of millions of posts generated per day on Twitter and other social media platforms.[8]

We leverage the abilities of DQE in a novel way to combat class imbalance in our dataset because the resulting class imbalance after our dataset curation was severe; our class distribution was 0.995% for age, 1.64% for ethnicity, 39.1% for gender, 11.7% for religion, and 46.6% for other (see Section V). This imbalance greatly affects the training process because of the resulting bias towards the Gender and Other classes, discounting Age and Ethnicity. In the existing literature, there are two ways to do this: undersampling or oversampling.

Undersampling reduces the majority class, which was not an attractive option, since the class with the lowest frequency had only 165 examples—too little for developing an effective classifier. Undersampling additionally wastes a large portion of our data that could provide a more representative picture of cyberbullying.

Instead, a few previous studies have used various oversampling techniques. Al-Garadi et al. [38], [39], used synthetic minority over-sampling technique (SMOTE), [40] used adaptive synthetic sampling (ADASYN). Both generate new points of the minority class by creating synthetic points along a line between minority data points. However, both oversampling methods have disadvantages: they can reduce the separation between classes, be sensitive to noise and outliers (which is perhaps more prevalent in short social media messages), be prone to overgeneralization, and are impractical for high dimensional data (BERT-based embedding methods produce vectors of 768 dimensions) [41]. Hence, this work presents a case for the use of DQE as an effective method to combat class imbalance, by gathering more natural data via semi-supervised learning, rather than generating synthetic examples.

[8]https://www.internetlivestats.com/twitter-statistics/

The purpose of DQE is to automatically identify the most representative words or features (called candidates) out of a textual data set. We initially hand-select a rudimentary input of seed queries that vaguely capture the main idea or theme that is being queried for (e.g. a good seed query for age might be "middle school"). We have modified the traditional DQE workflow to make it a semi-supervised and online algorithm consisting of four steps (see Figure 2):

1) Identify Target Space. We use a seed query set as a filter to identify the most representative part of the dataset, which enables the subsequent query generation to stay focused on a specific fine-grained class.

2) Rank Candidate Queries. Term-frequency inverse-document frequency (TF-IDF) weighting is applied to every word in each tweet of the target space, and the top 30 words become our candidate keywords for online query.

3) Online Query. This step is our modification to the DQE algorithm. After ranking candidates, we concatenate the top three candidates and expand the dataset with up to 5,000 more tweets with a call to the GetOldTweets3[9] library. If the top three candidates are not effective, a $\binom{5}{3}$ selection of the top five candidates was considered.

4) Update Seed Query and Reiterate. We repeat the previous three steps, using the top 30 candidate queries from this iteration as the seed queries for the next iteration. This process repeats until the difference between candidate significance values (TF-IDF weight) from the last iteration and the current iteration is within a predefined threshold value. This difference can be calculated by:

$$\frac{\sum\limits_{i \in C_t \setminus C_{t+1}} w_t(i) + \sum\limits_{j \in C_{t+1} \setminus C_t} w_{t+1}(j)}{\sum\limits_{k \in C_t} w_t(k) + \sum\limits_{l \in C_{t+1}} w_{t+1}(l)} \tag{7}$$

where t denotes the iteration, C_t denotes the set of candidates at iteration t, and w_t denotes the significance value of the candidate at iteration t. Our semi-supervised DQE method produces an impressively high quality dataset with few, if any,

[9]A library that uses JSON calls to scrape both new tweets and tweets older than a week old from Twitter. https://github.com/Mottl/GetOldTweets3

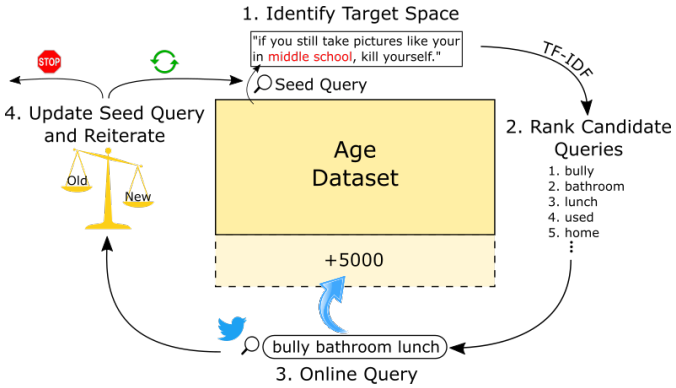


Fig. 2: Our Modified DQE Process

outliers upon a brief visual check. Our use and advancement of the DQE algorithm is novel because it is the first to our knowledge that seeks to (i) augment current datasets with the aim of solving class imbalance, (ii) integrate GetOldTweets3 for real-time updates and fresh data, (iii) use a combination of the first three candidates for real-time queries, (iv) run separate processes for each of the fine-grained classes, and (v) build off of already labeled data in a semi-supervised fashion.

V. Experiment

A. Experiment Data Generation

Based on our literature review, we were able to find 16 open source datasets from a wide variety of social media platforms and published by various authors, which is a surprisingly large number. Multiple papers have mentioned the relative scarcity of good cyberbullying datasets which can be compared [42]–[44]. We note that while the 16 datasets collectively contain a rather sufficient quantity of data, all of them are organized differently with varying fields and there has been no substantial work on compiling or maintaining a dataset that either combines all of the previous datasets or provides more than two or three classes of labels. Our work seeks to remedy this.

Taking all of the Twitter cyberbullying datasets, we ended up with the six datasets shown in Table I. The Bretschneider, Chatzakou, Waseem, and WISC datasets only provided Tweet IDs, so we used Twitter's API to retrieve the text content of the tweets. Given that many tweets have been taken down or removed since the publication of these datasets, we were able to retrieve 45.6%, 41.8%, 54.9%, and 51.4% of the tweets from those respective datasets. Because our research focuses on the fine-grained classification of cyberbullying tweets, we further classified the cyberbullying instances of these six datasets by hand and grouped tweets of the same class together (due to limited time and manpower, only the first 1500 tweets from the Chatzakou dataset and the first 4475 tweets from the Davidson dataset were further labeled and used). Our main contribution is utilizing a modified Dynamic Query Expansion, as mentioned in Section IV, to increase the number of samples of each class in a semi-supervised manner. We then randomly sampled 8000 tweets of Not CB, Age, Ethnicity, Gender, Religion, and Other to form our balanced dataset of size 48000.[10] For our experiments, we divided our train/test data by a 75:25 split.

The collective frequencies of fine-grained classes in the existing dataset before DQE highlight that ageism and racism may not be as represented in existing cyberbullying detection works. This is very concerning, and future studies may seek to see if previous research generalizes well to ageist or racist tweets. Our more representative dataset after using DQE is one of our major contributions in this work.[11]

Text Preprocessing: We performed very basic preprocessing on every tweet, stripping links, mentions (@username), the retweet flag "RT", and punctuation. Importantly, we did not remove hashtags or stop words. We believe that hashtags are still read as text, and therefore should be treated as such. We chose not to remove stopwords, as some of the commonly agreed upon stopword lists contain words that may be helpful context for the cyberbullying detection task, such as the use of "not" or the unusually frequent use of male or female pronouns. However, future studies may consider looking into the effect of stop words in cyberbullying detection.

B. Comparison Methods and Experiment Setup

Embedding methods are necessary to convert textual data into mathematical vectors that can serve as inputs to machine learning models. The goal is to generate representative vectors, where similar words have similar vectors. A popular comparative metric is the cosine similarity, which is the cosine of the angle between two vectors.

- **Bag of Words (BOW)**. In natural language processing, BOW is a simple feature extractor where the feature of each text is simply the frequency of each constituent word. The resulting embedding for our tweet dataset is a sparse matrix of size $\mathcal{N} \times \mathcal{V}$, where $\mathcal{N}$ is the number of tweets and $\mathcal{V}$ is the number of distinct vocabulary terms.
- **Term frequency-inverse document frequency (TF-IDF)**. TF-IDF is a feature extractor that operates similarly to BOW, by taking the frequencies of each word in a text and weighting them by how often that word appears in the whole document. We consider the document to be the entire training dataset.
- **word2vec**. Developed by Google, word2vec [45] applies a two-layer neural network to predict a word given its context (using the continuous bag of words (CBOW) implementation). This embedding model cannot handle out-of-vocabulary words, but remains as a popular choice for natural language processing tasks [40], [42], [43], as it tends to capture the relationships between words and phrases. We use the pre-trained 300-dimensional vectors trained from Google News.[12] The sentence/tweet embedding was taken by averaging the word vectors.

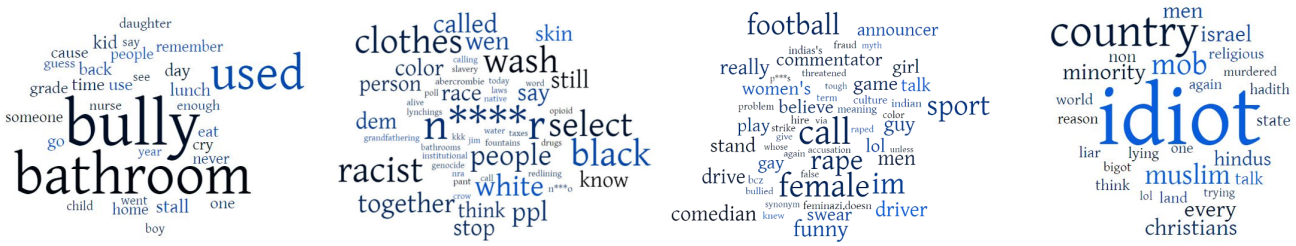
[10]Our experiment did not use the Not CB category as our detection method is meant as a downstream task after a traditional cyberbullying/not-cyberbullying classification, but we have included the not cyberbullying tweets for the community's use.

[11]For dataset, see footnote 7

[12]https://code.google.com/archive/p/word2vec/

TABLE I: Dataset Summary

Name	Total	CB	Not CB	Age	Ethnicity	Gender	Religion	Other
Agrawal [30]	16050	5963	10087	0	13	2841	1922	187
Brettschneider [33]	4475	183	4292	49	29	34	0	71
Chatzakou [29]	1500	1278	222	17	100	115	10	1036
Davidson [31]	205	181	24	5	27	121	1	27
Waseem [34], [35]	12899	8900	3999	0	86	3339	0	5475
WISC [32]	4095	1078	3024	94	17	39	0	921
Collective Before DQE	39224	17583	21648	165	272	6489	1933	7717
Collective After DQE	69767	50468	19299	10010	12730	10277	9367	8084



(a) Age (b) Ethnicity (c) Gender (d) Religion

Fig. 3: Word Clouds of Expanded Queries from DQE

- **GloVe**. Different from word2vec, GloVe is an unsupervised embedding model that leverages co-occurrences and aims to produce vectors whose dot products equal the logarithm of their co-occurrence probability. This enables GloVe to be better at encoding analogies. We use the pre-trained 300 dimensional vectors trained from 840 billion tokens in Common Crawl [46].
- **fastText.** Developed by Facebook, fastText is built off of word2vec, but is capable of handling out-of-vocabulary words by encoding words as its n-gram representation. We used the English 300-dimension pre-trained fastText vectors, which had been trained on texts from Common Crawl and Wikipedia [47].
- **BERT.** The above methods are not context-aware and cannot capture multiple definitions of a particular term. Google research has developed bidirectional encoder representations from transformers (BERT), which uses transformer networks trained on predicting occluded words in a sentence to learn context-specific word embeddings. However, these models are known to be large, and the output vectors have 768 dimensions. We used the Huggingface transformers library for their BERT implementation [48]. The sentence/tweet embedding was taken from the first hidden state of the head [CLS] token.
- **DistilBERT.** Another transformer-based text embedding method, DistilBERT is a compact version of BERT, with "40% less parameters than bert-base-uncased, runs 60% faster while preserving over 95% of Bert's performances as measured on the GLUE language understanding benchmark." We used the Huggingface transformers library for their DistilBERT implementation [48].
- **Sentence BERT (SBERT).** All the above embedding methods aren't designed to create sentence/tweet embeddings that can be fed into a classifier model. Sentence BERT fine-tunes a BERT model for the specific task of sentence embeddings tested on the semantic textual similarity (STS) benchmark using a siamese network structure. This is one of the SOTA embedding methods in NLP. We used UPKLab's sentence transformer library with the pre-trained bert-base-nli-stsb-mean-tokens model [49].

To test our SOSNet model, we experimented with six other traditional machine learning classifiers that have precedent in previous cyberbullying detection studies, as discussed in [43]. These include **logistic regression (LR), Naïve Bayes (NB), k-nearest neighbors (KNN), support vector machine (SVM), XGBoost (XGB), and multi-layer perceptron (MLP).** All of these methods were implemented with sklearn.

With our comparison methods in place, we went through five steps for each tweet embedding+classifier model combination: word embedding generation,[13] feature extraction, normalization, running/tuning with 5-fold stratified cross-validation (emulating the structure used by [42]), and statistical analysis.

We repeat this process for a randomly sampled, downscaled version of the dataset that is only 10% of the size of the original dataset to test the scalability of these models. We ran our experiments through Google Colaboratory on a Tesla P100 GPU, with 25 GB of RAM and 147 GB of disk space.

For comparing two models, we used sklearn to compute the confusion matrices and evaluation metrics (accuracy and F1 score) and conduct McNemar's tests as popularized in [17] due to the inability to assume normality or independency for a matched pair t-test. The distribution cannot be assumed to be

[13]For SBERT, BOW, and TF-IDF, no word embeddings were generated; the output of these models were already features. We used the methods discussed in Section V-B to extract features for a tweet/sentence embedding, and used sklearn's StandardScaler utility to normalize the resulting vectors (and MinMaxScaler for the Naïve Bayes classifier, except for TF-IDF and BOW, where MaxAbsScaler was used instead).

normal since the number of samples (times the model is run) is less than 30, so therefore the Central Limit Theorem (CLT) does not apply. Also, because some of the data points were derived from DQE, proper independency may be contested. A McNemar's test is computed on a 2x2 contingency table of whether the predicted label matched the ground truth for the two models being compared.

C. Experimental Design

The purpose of our experiment is to investigate the efficacy of different tweet embedding and classifier model combinations for differentiating the five classes/targets of cyberbullying (age, ethnicity, gender, religion, other) in an automatic cyberbullying detection system across two magnitudes of data. Our independent variable was the combination of tweet embedding and classifier model used, and our dependent variable was the resulting performance of the combination, measured by the evaluation metrics that we selected: accuracy and F1 score. We calculate the accuracy as a baseline measure because it is the most intuitive evaluation metric and the classes are balanced because of DQE. We also calculate the F1 score (the harmonic mean between precision and recall) because of its popularity among machine learning model comparisons and because we care both about precision and recall: precision is the rate at which the classifier correctly identifies the type of attack out of all the tweets it predicts as that type of attack, while recall is the rate at which the classifier correctly identifies the type of attack when given tweets of that type of attack. Our null hypothesis is that SBERT+SOSNet will have no statistically significant difference between the evaluation metrics across the levels of IV. Our alternative hypothesis is that SBERT+SOSNet will have a statistically significant difference between the evaluation metrics across the levels of IV.

D. Fine-Grained Cyberbullying Classification Results

TABLE II: Test Accuracies—40,000 Tweets

Accuracy	LR	NB	KNN	SVM	XGB	MLP
SBERT	0.8982	0.8063	0.8555	0.9267	0.9151	0.9148
BERT	0.8791	0.6938	0.7759	0.8977	0.8786	0.8873
DistilBERT	0.9033	0.7280	0.8316	0.9187	0.9058	0.9050
GloVe	0.8804	0.6828	0.7726	0.9225	0.9167	0.9154
W2V	0.8806	0.6898	0.7672	0.9210	0.9212	0.9083
FastText	0.8671	0.5994	0.7038	0.9077	0.9004	0.9011
TF-IDF	0.8514	0.8265	0.3086	0.8095	0.9378	0.8375
BOW	0.8728	0.8041	0.5364	0.8419	0.9438	0.8783

TABLE III: Test F1 Scores—40,000 Tweets

F1 Score	LR	NB	KNN	SVM	XGB	MLP
SBERT	0.8981	0.8066	0.8488	0.9272	0.9157	0.9149
BERT	0.8792	0.6944	0.7740	0.8981	0.8791	0.8877
DistilBERT	0.9033	0.7287	0.8302	0.9190	0.9061	0.9051
GloVe	0.8799	0.6743	0.7421	0.9230	0.9171	0.9153
W2V	0.8809	0.6811	0.7328	0.9215	0.9217	0.9082
FastText	0.8672	0.5920	0.6792	0.9080	0.9007	0.9009
TF-IDF	0.8528	0.8157	0.2797	0.8116	0.9386	0.8375
BOW	0.8747	0.7876	0.5281	0.8448	0.9444	0.8786

40,000 Tweets: Our experiments show that BOW+XGBoost and TF-IDF+XGBoost outperform every other method on the full dataset in both accuracy and F1 scores. BOW and TF-IDF are the simplest tweet embeddings that we tested: they simply count word frequencies, and TF-IDF weights those frequencies. At its core, XGBoost also uses one of the most basic models: decision trees. Our results demonstrate that simple models may not only be the most interpretable, but may also prove to be more effective than complex models. Based on the favorable outcomes of the BOW and TF-IDF classifiers and the more decision-based quality of XGBoost, we suspect that vocabulary or combinations of vocabulary may be the best discriminating factor to tell between cyberbullying targets in this fine-grained classification task. Part of this may also be attributed to the DQE method for balancing classes, as the most representative keywords of each class were used to retrieve new samples, which may decrease the diversity of vocabulary in the dataset.

Apart from the simple models, SBERT stood out as the generally most effective embedding method across each classifier model. SVM stood out as the generally most effective classifier model across each embedding method. As SBERT is one of the current SOTA sentence embedding methods for natural language inference and the semantic textual similarity benchmark, it is unsurprising that it outperforms other embedding methods. Further advancements in natural language understanding embeddings can be expected to generally improve current automatic cyberbullying detection.

Of the transformer-based methods, DistilBERT outperforms BERT, to much surprise. Of the word embedding-based methods, GloVe and word2vec are similar, and both outperform fastText, even though fastText is the only method capable of embedding out-of-vocabulary words. GloVe and word2vec are comparable to SBERT as well, as their results have no statistically significant difference at a 0.05 significance level (see Table IV).

TABLE IV: Select McNemar's Test Calculations[14]

Model A	Model B	χ^2	p-value	$p < .05$
BOW+XGBoost	TF-IDF+XGBoost	3130	0.0	Yes
BOW+XGBoost	SBERT+SVM	250	2.87E-56	Yes
TF-IDF+XGBoost	SBERT+SVM	3930	0.0	Yes
SBERT+SVM	SBERT+XGBoost	632	1.8E-139	Yes
SBERT+SVM	SBERT+MLP	27.5	1.57E-07	Yes
SBERT+MLP	SBERT+XGBoost	393	1.62E-87	Yes
SBERT+SVM	DistilBERT+SVM	12.4	4.40E-04	Yes
SBERT+SVM	GloVe+SVM	2.96	0.0854	No
GloVe+SVM	fastText+SVM	6910	0.0	Yes
DistilBERT+SVM	BERT+SVM	70.9	3.76E-17	Yes
GloVe+SVM	word2vec+SVM	0.450	0.450	No

TABLE V: Test Accuracies—4,000 Tweets

Accuracy	LR	NB	KNN	SVM	XGB	MLP	SOSNet
SBERT	0.8900	0.8420	0.8240	0.9200	0.8950	0.9160	0.9270
BERT	0.8330	0.6820	0.7060	0.8480	0.8280	0.8600	0.8580
DistilBERT	0.8900	0.7320	0.7990	0.8840	0.8830	0.9030	0.8890
GloVe	0.8520	0.7000	0.7320	0.9060	0.8970	0.8870	0.8830
W2V	0.8280	0.7140	0.7230	0.9050	0.8930	0.8860	0.8710
FastText	0.8280	0.6090	0.6740	0.8710	0.8710	0.8680	0.8850
TF-IDF	0.8330	0.7890	0.2350	0.6460	0.9030	0.7340	0.7170
BOW	0.8350	0.7750	0.4180	0.6580	0.9180	0.7600	0.7670

[14]These are statistics of interest; the rest are available upon request.

TABLE VI: Test F1 Scores—4,000 Tweets

F1 Score	LR	NB	KNN	SVM	XGB	MLP	SOSNet
SBERT	0.8876	0.8403	0.8103	0.9190	0.8936	0.9142	0.9258
BERT	0.8305	0.6803	0.7034	0.8471	0.8264	0.8586	0.8552
DistilBERT	0.8878	0.7287	0.7953	0.8815	0.8805	0.9009	0.8876
GloVe	0.8480	0.6887	0.6916	0.9050	0.8954	0.8829	0.8825
W2V	0.8221	0.7038	0.6717	0.9028	0.8908	0.8821	0.8697
FastText	0.8245	0.6005	0.6427	0.8697	0.8690	0.8645	0.8833
TF-IDF	0.9298	0.7612	0.1559	0.6602	0.9028	0.7327	0.7113
BOW	0.8362	0.7408	0.3975	0.6721	0.9177	0.7636	0.7598

TABLE VII: Select McNemar's Test Calculations[15]

Model A	Model B	χ^2	p-value	$p < .05$
SBERT+SOSNet	SBERT+SVM	1.140	0.286	No
SBERT+SOSNet	SBERT+XGBoost	48.95	2.63E-12	Yes
SBERT+SOSNet	BOW+XGBoost	66.77	3.06E-16	Yes
SBERT+SOSNet	DistilBERT+SOSNet	15.36	8.88E-05	Yes
GloVe+SOSNet	fastText+SOSNet	0.03846	0.845	No

4,000 Tweets: Our results show that on the downsized version of the data, the SBERT+SOSNet combination produced the best accuracy and F1 score. The TF-IDF and BOW models generally show the largest decline in performance across both evaluation metrics. TF-IDF+XGBoost and BOW+XGBoost, however, still remain as strong models. SBERT+SVM has the second highest set of evaluation metrics, and a McNemar's test with SBERT+SOSNet shows that there is no statistically significant difference to the two models at the 0.05 significance level. This shows that the SOSNet classifier can match the performance of top-of-the-line SVM models. A McNemar's test between SBERT+SOSNet and BOW+XGBoost shows a statistically significant difference at the 0.05 significance level, indicating that SBERT+SOSNet provides superior results, at least at this scale of data.

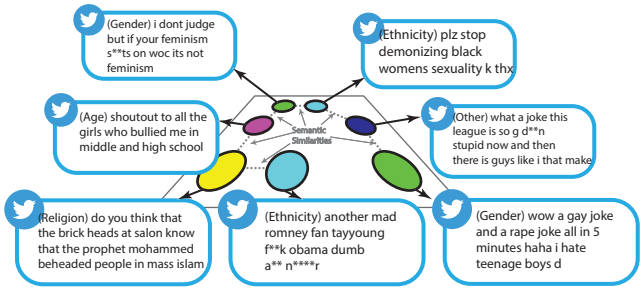


Fig. 4: Example of Cyberbullying Detection from SOSNet

Case Study: Of the 1000 tweets that were tested, the SBERT+SOSNet combination misclassified a total of 73 tweets. Of the 73 tweets, only 14 tweets came from DQE-generated data (19.2%). This demonstrates the robustness of the DQE process as compared to the manual gathering of data. On analysis of the confusion matrix produced from our experiments, across the board, the most confused classes in this task were gender and other, and these misclassifications account for a majority of the error. Figure 4 shows a sample cyberbullying detection result from our constructed semantic similarity graph.

Annotator error from the six initial datasets may account for much of the models' error, as many of the misclassifications should be classified as the machine learning model predicts.

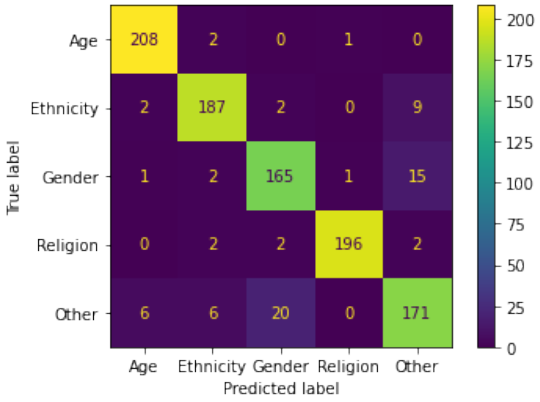


Fig. 5: Confusion Matrix for SBERT+SOSNet

For example, this tweet's ground truth was labeled as religion: "my 2282 says that you can replace the testimony of one man with that of two women," but the model predicts gender (rightly so). This indicates that the reported results may be underestimates of these models' real-world accuracies.

VI. Conclusion

The purpose of our paper is to establish an effective fine-grained cyberbullying classifier to root out hateful tweets before an escalation into cyberbullying occurs. First, we demonstrate that Dynamic Query Expansion effectively combats class imbalance, and we recommend its use for other applications in social media data mining and natural language processing. Second, we pioneer fine-grained cyberbullying detection, conducting meticulous experiments with a multitude of embedding method+classifier model combinations, and show similar progress to previous binary cyberbullying classification studies. Finally, our proposed Graph Convolutional Network framework SOSNet capitalizes on inherent semantic connections between tweets, and our results show that this approach matches or exceeds the performance of traditional classifiers in this domain. Our research represents a step forward for establishing an active anti-cyberbullying presence in social media, and a step forward towards a future without cyberbullying.

References

[1] E. Englander, E. Donnerstein, R. Kowalski, C. A. Lin, and K. Parti, "Defining cyberbullying," *Pediatrics*, vol. 140, no. Supplement 2, pp. S148–S151, 2017.

[2] L. Johnson, "Counselors and cyberbullying: Guidelines for prevention, intervention, and counseling," *Retrieved January*, vol. 7, p. 2015, 2011.

[3] J. Salminen, M. Hopf, S. A. Chowdhury, S.-g. Jung, H. Almerekhi, and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Human-centric Computing and Information Sciences*, vol. 10, no. 1, p. 1, 2020.

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

[5] M. Dadvar and F. De Jong, "Cyberbullying detection: a step toward a safer internet yard," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 121–126.

[15]These are statistics of interest; the rest are available upon request.

[6] M. Dadvar, F. d. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent, 2012.

[7] V. S. Chavan and S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2015, pp. 2354–2358.

[8] A. Squicciarini, S. Rajtmajer, Y. Liu, and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 2015, pp. 280–285.

[9] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Multi-task learning for spatio-temporal event forecasting," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1503–1512.

[10] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Multi-resolution spatial event forecasting in social media," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 2016, pp. 689–698.

[11] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 851–860.

[12] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic crime prediction using events extracted from twitter posts," in *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, 2012, pp. 231–238.

[13] L. Zhao, J. Chen, F. Chen, W. Wang, C.-T. Lu, and N. Ramakrishnan, "Simnest: Social media nested epidemic simulation via online semi-supervised deep learning," in *2015 IEEE International Conference on Data Mining*. IEEE, 2015, pp. 639–648.

[14] J. Piskorski, H. Tanev, and A. Balahur, "Exploiting twitter for border security-related intelligence gathering," in *2013 European Intelligence and Security Informatics Conference*. IEEE, 2013, pp. 239–246.

[15] F. Atefeh and W. Khreich, "A survey of techniques for event detection in twitter," *Computational Intelligence*, vol. 31, no. 1, pp. 132–164, 2015.

[16] A. Ritter, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1104–1112.

[17] H. Becker, D. Iter, M. Naaman, and L. Gravano, "Identifying content for planned events across social media sites," in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 533–542.

[18] H. Tanev, M. Ehrmann, J. Piskorski, and V. Zavarella, "Enhancing event descriptions through twitter mining." in *ICWSM*. Citeseer, 2012.

[19] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," in *ICWSM*, 2011.

[20] X. Zhou and L. Chen, "Event detection over twitter social media streams," *The VLDB journal*, vol. 23, no. 3, pp. 381–400, 2014.

[21] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz *et al.*, "'beating the news' with embers: forecasting civil unrest using open source indicators," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1799–1808.

[22] R. P. Khandpur, T. Ji, Y. Ning, L. Zhao, C.-T. Lu, E. R. Smith, C. Adams, and N. Ramakrishnan, "Determining relative airport threats from news and social media," in *Twenty-Ninth IAAI Conference*, 2017.

[23] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[24] M. Z. Moricz, L. M. Monier, and J.-C. Michelou, "Dynamic query expansion," Jun. 25 2002, uS Patent 6,411,950.

[25] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *Acm Computing Surveys (CSUR)*, vol. 44, no. 1, pp. 1–50, 2012.

[26] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Dynamic theme tracking in twitter," in *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2015, pp. 561–570.

[27] L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, and N. Ramakrishnan, "Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling," *PloS one*, vol. 9, no. 10, p. e110206, 2014.

[28] L. Zhao, J. Chen, F. Chen, F. Jin, W. Wang, C.-T. Lu, and N. Ramakrishnan, "Online flu epidemiological deep modeling on disease contact network," *GeoInformatica*, vol. 24, no. 2, pp. 443–475, 2020.

[29] D. Chatzakou, I. Leontiadis, J. Blackburn, E. D. Cristofaro, G. Stringhini, A. Vakali, and N. Kourtellis, "Detecting cyberbullying and cyberaggression in social media," *ACM Transactions on the Web (TWEB)*, vol. 13, no. 3, pp. 1–51, 2019.

[30] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European Conference on Information Retrieval*. Springer, 2018, pp. 141–153.

[31] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *arXiv preprint arXiv:1703.04009*, 2017.

[32] 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*, 2012, pp. 656–666.

[33] U. Bretschneider, T. Wöhner, and R. Peters, "Detecting online harassment in social networks," in *ICIS*, 2014.

[34] 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.

[35] 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.

[36] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, 2011.

[37] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[38] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[39] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges," *IEEE Access*, vol. 7, pp. 70 701–70 718, 2019.

[40] M. Al-Hashedi, L.-K. Soon, and H.-N. Goh, "Cyberbullying detection using deep learning and word embeddings: An empirical study," in *Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems*, 2019, pp. 17–21.

[41] S. J. Dattagupta, "A performance comparison of oversampling methods for data generation in imbalanced learning tasks," Ph.D. dissertation, NOVA University Lisbon, 2018.

[42] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *European semantic web conference*. Springer, 2018, pp. 745–760.

[43] H. Rosa, N. Salgado Pereira, R. Ribeiro, P. Ferreira, J. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. Veiga Simão, and I. Trancoso, "Automatic cyberbullying detection: A systematic review," *Computers in Human Behavior*, vol. 93, pp. 333–345, 04 2019.

[44] C. Emmery, B. Verhoeven, G. De Pauw, G. Jacobs, C. Van Hee, E. Lefever, B. Desmet, V. Hoste, and W. Daelemans, "Current limitations in cyberbullying detection: on evaluation criteria, reproducibility, and data scarcity," *arXiv preprint arXiv:1910.11922*, 2019.

[45] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

[46] 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.

[47] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," *arXiv preprint arXiv:1802.06893*, 2018.

[48] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, pp. arXiv–1910, 2019.

[49] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.