Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1761–1770
Marseille, 20-25 June 2022

Data Expansion Using WordNet-based Semantic Expansion and Word Disambiguation for Cyberbullying Detection

Md Saroar Jahan[⋆], Djamila Romaissa Beddiar[⋆], Mourad Oussalah[⋆], Muhidin Mohamed[◇]
[⋆]University of Oulu, CMVS, BP 4500, 90014, Finland
[◇] Operations and Information Management, Aston University, B4 7ET, UK
{Md.Jahan, Mourad.Oussalah, Djamila.Beddiar}@oulu.fi, m.mohamed10@aston.ac.uk

Abstract

Automatic identification of cyberbullying from textual content is known to be a challenging task. The challenges arise from the inherent structure of cyberbullying and the lack of labeled large-scale corpus, enabling efficient machine-learning-based tools including neural networks. This paper advocates a data augmentation-based approach that could enhance the automatic detection of cyberbullying in social media texts. We use both word sense disambiguation and synonymy relation in WordNet lexical database to generate coherent equivalent utterances of cyberbullying input data. The disambiguation and semantic expansion are intended to overcome the inherent limitations of social media posts, such as an abundance of unstructured constructs and limited semantic content. Besides, to test the feasibility, a novel protocol has been employed to collect cyberbullying traces data from AskFm forum, where about a 10K-size dataset has been manually labeled. Next, the problem of cyberbullying identification is viewed as a binary classification problem using an elaborated data augmentation strategy and an appropriate classifier. For the latter, a Convolutional Neural Network (CNN) architecture with FastText and BERT was put forward, whose results were compared against commonly employed Naïve Bayes (NB) and Logistic Regression (LR) classifiers with and without data augmentation. The research outcomes were promising and yielded almost 98.4% of classifier accuracy, an improvement of more than 4% over baseline results.

Keywords: cyberbullying detection, social media, disambiguation, dataset expansion, WordNet

1. INTRODUCTION

The multiplication of cases associated with hate speech (HS) and cyberbullying nowadays, especially on social media platforms (Gong et al., 2014), raises concerns about the efficiency of adopted measures. This becomes a challenge for policymakers, educators, sociologists, civil society actors, and researchers. Indeed, the advances in Web 2.0 technology, which led to the emergence of anonymously user-generated content, have witnessed the proliferation of cyberbullying cases in public forums and political discourse. Research in cyberbullying has been mainly multidisciplinary by nature, where several definitions have been promoted in the literature depending on the research focus. Possibly, the definition that best accommodates our work in this paper is that reported in (Rosa et al., 2019) where cyberbullying is associated with an individual's or group's repeated intentional aggressive behavior towards other peers with the intent of harming them by sending offensive content or engaging in other forms of social aggression through the use of digital technologies. Another related cyberbullying definition is provided by (Chen, 2011) as an 'aggressive and intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who can not easily defend him or herself'. However, there is a subtle difference between cyberbullying and HS: 'hate speech is more general and not necessarily focused on a specific person' (Fortuna and Nunes, 2018). Besides, (Nockleby, 2000) stated that 'hate speech disparages a person or group based on some characteristics such as race, color, and ethnicity'.

Many researchers have started to explore automatic procedures for large-scale social media monitoring and early detection of harmful events, including cyberbullying. Current solutions adopted by large-scale companies (e.g., news agencies, Facebook, Twitter, public forum associations, Reddit) rely mainly on user reporting of objectionable content. However, this is trivially time-consuming, labor-intensive, and it is neither sustainable nor scalable (Salawu et al., 2017). An effort to automatically identify HS or cyberbullying from textual content has mainly focused on the application of machine learning approaches, which view the identification process as a pure classification. Although, this tends to work well when trained on a large-scale dataset (Bello et al., 2017), it falls short when this is not the case. Besides, the collection and annotation of large-scale cyberbullying datasets are very challenging (Zhang and Luo, 2019) due to inherent subjectivity in comprehending cyberbullying cases, especially in cases of multiple fine-grained hate-speech categories, i.e., insults, threats (Van Hee et al., 2015). The widespread use of shorthand English and abbreviations on social media also poses another challenge in identifying the meaning of cyberbullying texts(Kumar and Sachdeva, 2020). In such cases, an effective solution is to produce large datasets by expanding a small well-annotated dataset. Data augmentation is an effective technique to increase both the amount and diversity of low-resource language datasets (Bello et al., 2017). Developing sound data augmentation approaches suitable for leveraging the lack of good quality cyberbullying datasets is among the main motives of the current work, aiming to contribute to the lack of scalability and to overcome significant bias observed in non-hate speech detection.

Benefiting from the usage of natural language processing techniques (NLP) in this challenging field, the paper posits our main contributions as follows:

- Three different synonym-based augmentation schemes have been put forward. Two key princi-

ples guide this augmentation: i) the word sense disambiguation (WSD) is implemented through Lesk-algorithm (Lesk, 1986), which is employed to capture the true sense of each word of a given statement, and ii) the use of WordNet lexical dataset (Fellbaum, 2010) to fetch the corresponding senses that will be used to generate new equivalent hate-speech statements.

- We compared our proposed data augmentation approach against the state-of-art Mixup, a newly proposed data augmentation method through linearly interpolating inputs and modeling targets of random samples, which was introduced for natural language processing in (Guo et al., 2019).
- We developed a new python library for data augmentation, which is the end product of our experiment, and released under an open-sourced license to the research community [1].
- We released a new cyberbullying dataset and considered the subtle differences between common hate speech and cyberbullying during annotation. After that, an original setup was conducted to test the developed methodology using both a newly collected cyberbullying dataset from the AskFm forum and a publicly available dataset (FormSpring dataset). In addition, six newly extended datasets for cyberbullying task are released publicly.
- We exploited FastText as a classifier and compared its performance with FastText as a word embedding with CNN classifier. Besides, we compared the BERT and CNN classifier results with other machine learning baseline classifiers, namely, Linear Regression and Naive Bayesian classifiers.

The paper is structured as follows: Section 2 presents a brief review of related work in the field. Section 3 highlights the overall methodology. We discuss the experimental results in Section 4, and finally, we conclude the paper in Section 5 and outline some future research directions.

2. RELATED WORK

The need for implementing automatic cyberbullying detection mechanisms is crucial and essential to capture, track, and prevent the incidence of bullying trace, especially in online platforms. In this respect, starting from the pioneering work of (Warner and Hirschberg, 2012), the use of machine-learning-based classifiers for detecting abusive language became popular within the information processing research community. For instance, (Nobata et al., 2016) combined pre-defined language elements and word embedding to train a regression model. Nevertheless, the current machine-learning-based approaches are obstructed by the challenges associated with the definition of hate speech discourse. Indeed, the presence of an insult, for example, does not necessarily entail a hate speech post. Similarly, the constant evolution of HS corpus and the variety of expressions therein bring an extra limitation to the scope of the training samples. Besides, mainly due to the small scale of hate-speech dataset as compared to non-hate speech dataset, (Burnap and Williams, 2015) reported that many of the existing hate-speech detection approaches are largely biased towards detecting content that is non-hate as opposed to detecting and discriminating real hateful content. Furthermore, some significant work has been done for automated cyberbullying detection using social network datasets such as AskFm datasets (Foong and Oussalah, 2017). (Dinakar et al., 2012) conducted hate-speech text classification experiments on YouTube data, while an annotated cyberbullying dataset and a fine-grained classification are put forward in (Van Hee et al., 2015). Recently, many works focused on deep learning-based models to identify the aggressive language in social media texts. For instance, (Agrawal and Awekar, 2018) investigated how learning-based models can capture more dispersed features on various platforms and topics. (Bu and Cho, 2018) provided a hybrid deep learning model that combines CNN and Long-term Recurrent Convolutional Networks (LRCN) to detect cyberbullying in Social Networking Service (SNS) comments. A character-level CNN model with shortcuts was proposed by (Lu et al., 2020). (Rosa et al., 2018) compared three different deep learning approaches, trained from three different sources for multiple category textual cyberbullying detection. On the other hand, in the absence of large-scale labeled corpus(Jahan and Oussalah, 2021), an intuitive approach is to seek an appropriate data augmentation strategy. Often, the whole meaning of the sentence is radically changed when making a slight change of a word. Many approaches rely on word replacements to transform the sentences and expand the dataset accordingly, as in (Kobayashi, 2018; Sahin and Steedman, 2018). However, word replacement based techniques are not always efficient due to the smallness of the vocabulary, lack or unsuitability of synonym terms, and difficulty in maintaining the underlined context. (Zhang et al., 2017) has introduced an effective augmentation method called Mixup used in image classification models and showed superior performance. An adaptation of Mixup to sentence classification was presented by (Guo et al., 2019) with two strategies: one performs interpolation on word embedding and another one on sentence embedding.

On the other hand, several works have been reported in the context of word sense disambiguation (WSD) using the so-called Lesk algorithm or extended lesk algorithm with Wordnet, and their variants (Ekedahl and Golub, 2004; Naskar and Bandyopadhyay, 2007). Typically, WSD automatically identifies the meaning of a given target word in its associated context. It has drawn much interest in the last decade, and many improved results are being obtained. For instance, it has been reported that the Extended-WordNet based word-sense disambiguation for noun, verb, and adjective categories achieved a precision of 85.9% (Naskar and Bandyopadhyay, 2007). Since its emergence, many studies are conducted to tackle the problem of WSD. For instance, (Gutiérrez et al., 2017; Chaplot and Salakhutdinov, 2018) developed new systems for WSD. Roughly speaking, (Gutiérrez et al., 2017) applied the PageRank

[1] https://pypi.org/project/nlp-augment/

algorithm using different semantic dimensions (resources) and adding word-sense frequency knowledge to improve the results. Likewise, (Chaplot and Salakhutdinov, 2018) proposed a novel knowledge-based WSD algorithm for the all-word WSD task, which utilizes the whole document as the context for a word, rather than just the current sentence used by most WSD systems. This motivates our core idea of using WSD for cyberbullying dataset expansion, which, as far as we know, has not been experimented with a complete dataset or corpus. However, a work by (Beddiar et al., 2021) has performed data augmentation for hate speech corpus using Back Translation and Paraphrasing and reported 99.7% accuracy. This motivates us to experiment with only synonym replacement data augmentation and more contextual sense-based augmentation techniques. Our work yields a new expanded dataset, which, in turn, provides useful insights to handle the imbalanced class created by small instances of labeled cyberbullying cases as compared to non-cyber bullying cases. Furthermore, several works have been done, including using synonyms and more targeted synonym replacement example, POS tag (Sun and He, 2020; Jungiewicz and Smywiński-Pohl, 2019); however, the differences and benefits between using simple synonym replacement or more targeted synonym replacement have been overlooked in previous research. This work will try to answer whether a more specific contextual synonym replacement works better for a machine learning model than a more general synonym replacement technique.

3. METHODOLOGY

Our experiment methodology includes a three-fold process. First, a newly collected dataset from AskFm social network website and the publicly FormSpring dataset *(Reynolds et al., 2011)* were introduced. Next, data augmentation is performed using the Wordnet-based sense disambiguation technique and Lesk-algorithm *(Lesk, 1986)*. Our proposed techniques are compared to the state-of-the-art data augmentation method (Mixup). Then, a classification approach involving BERT, CNN, NB, LR, and FastText models was devised. The results are contrasted and the data augmentation process has been duly evaluated for both AskFm and FormSpring datasets.

3.1. Datasets

AskFm Dataset: The first original dataset that we have used in this paper is collected from AskFm website [2]. It is primarily used for asking questions either publicly or anonymously and then getting answers from other users. To collect each user's questions and answers, we have crawled each of the profiles using Python web crawler library, Beautiful Soup[3]. Questions and answers associated with each user profile are saved in a CSV file. Question-answer pairs are only extracted if they contain cyberbullying swear words, which were filtered by a string matching technique. In total, we crawled 3720 user profiles and over 400,000 question-answer pairs. Public proxy servers located in the UK, CA, and the USA were employed to retrieve English written posts. Applying insult/swear words string matching reduced the data to 10k unique posts, containing at least one insult/swear word either in the question or in the answer parts.

We have manually labeled the resulting 10k AskFM dataset. Labeling involves identifying whether each sentence contains cyberbullying or not. A critical hypothesis developed based on some social science and psychiatry findings, that cyberbullying cases must include both insult/swear wording and a second person/person's name (Patchin and Hinduja, 2006). If there is no second person/person's name available, it may not be considered cyberbullying other than general HS. However, this determination is not that simple. For example, "This is bad, but John Doe is lucky" includes both Insult word and Person name; however, it is not an HS or cyberbullying case as the relationship between the two is not established. Nevertheless, "John Doe is not bad," contains both person name, Insult word, and a correlation between the two has been established. Still, it is not an HS or cyberbullying case due to the negation of words. Similarly, 'John Doe is not a good person,' does not contain Insult words but is considered as both HS and cyberbullying. In the sentence, 'Asylum seekers are dirty' has an Insult word and target; it falls into HS but is not categorized as cyberbullying because its target is not a person / second person. Besides, HS could work differently if multiple sentences are put together. For example, "John doe working hard. Ugly" is an HS, Cyberbullying and Offensive cases even though the second sentence "Ugly" contains only an Insult word without any second-person/Person entity. These examples show the requirements mentioned above for HS, cyberbullying, and other cases are the necessary conditions; though, it is not compulsory due to the variety of natural language modifiers expressing negation and opposition.

Two independent labelers (who have knowledge in this field and completed a master's thesis on cyberbullying detection and NLP) have been employed separately for annotation for avoiding bias. While a third one (a senior research fellow and completed his Ph.D. in this field) is called upon whenever a disagreement between the two arises (a total of 141 sentences were disagreed). We attribute the label '1' to a sentence if it is associated with cyberbullying and '0' otherwise. Once labeled, 21.3% of the dataset was identified as cyberbullying and the rest as non-cyberbullying. The details of the dataset collection source code and datasets will be released for the community on this GitHub page [4].

FormSpring Dataset The second dataset that we have used was collected from fromspring.me *(Reynolds et al., 2011)*, which is publicly accessible. The data represent 50 IDs from formspring.me that were crawled in the Summer of 2010. For each ID, the profile information and each post (question and answer) were extracted. Each post was loaded into Amazon's Mechanical Turk and labeled by three workers for cyberbullying content. The same labeling mechanism as before is used here, as illustrated in Table 1. The FormSpring dataset contains 12k posts in which 7%

[2] https://ask.fm/ (accessed Oct 03, 2019)

[3] https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (accessed Oct 03, 2019)

[4] https://github.com/saroarjahan/expansion_of_cyberbyllying_datasets/

contains cyberbullying.

Table 1: Labelling example from original dataset.

Questions	Label
how to tell if a guy is gay if they super straight	0
You are gay	1

3.1.1. Expanded Datasets

To effectively capture all intended meanings of a given cyberbullying text and hence improve its detection, the experimental datasets were artificially enriched. In other words, the original base datasets were extended with the aim of expanding the semantic space of each initial cyberbullying sentence in the datasets. For this purpose, we have proposed three possible methods briefly summarized below and detailed in the pseudo-code "Algorithm 1".

Algorithm 1: Generate new list of sentences for expanded datasets:

```
(i) INPUT: Load each Sentence;
(ii) Initialize the expanded dataset with the original sentence;
(iii) Perform word Tokenization and make a list of words;
(iv) Remove stop words;
(v) for each word do
    switch method do
        case method1 do
            disambiguate with Lesk and find sense specific Synset;
            if sense specific Synset has Synonyms then
                for each synonym do
                    Replace sentence word with synonym;
                    Generate new sentence;
                    Append the new sentence to the expanded dataset;
                end
            end
        end
        case method2 do
            detect POS tag and find sense specific Synset;
            if sense specific Synset has Synonyms then
                for each synonym do
                    if synonym POS tag is equal to word POS tag then
                        Replace sentence word with synonym;
                        Generate new sentence;
                        Append the new sentence to the expanded
                          dataset;
                    end
                end
            end
        end
        case method3 do
            if Synset has Synonyms then
                for each synonym do
                    Replace sentence word with synonym;
                    Generate new sentence;
                    Append the new sentence to the expanded dataset;
                end
            end
        end
        otherwise do
            Error: No such method
        end
    end
end
```

Method 1: We applied word-sense disambiguation to each word of the input sentence, after the preprocessing stage that removes stopwords and other uncommon characters. The synonymy relation was used to extract the list of senses for each word. Next, to find out which of these senses better fit the context of the sentence, Lesk algorithm *(Lesk, 1986)* was employed. The original version of this algorithm disambiguates words in short sentences. For that, the gloss of the word to disambiguate (dictionary of its senses) is compared to glosses of other words of the sentence. Then, the sense that shares the most significant number of common words with the glosses of other words of the phrase is chosen and assigned to the target word. Fig 1 illustrates an example of the application of this algorithm to the sentence "He is gay" and its newly generated sentences.

Method 2: We apply Part-of-Speech (PoS) Tagging to each sentence, which is later used to extract all meanings (synsets) and synonyms that correspond to that word #PoS combination. This approach could widely expand the semantic space over the previously mentioned data augmentation approach (method 1), as one word could have multiple meanings in the same part of speech.

Method 3: We extract all possible meanings (synsets) of every complete word (after preprocessing), and then we retrieve the synonyms associated with every possible meaning. This significantly expands the semantic space of each sentence compared to the first two methods. We are considering here all possible meanings (including every PoS that this word may belong to) as well as the similar words of each meaning regardless of the coherence of the corresponding context.

To apply the proposed methodology, we have written a python script that generates extended datasets. This is achieved by following the above-described methods for each of the original datasets. Table 2 compares the size of the original and expanded datasets. Examples of some generated sentences are provided in Tab. 3. Notably, one can perceive the intuition and quality of the generated new sentences, where the algorithm successfully created semantically similar sentences. However, some generated sentences failed to hold the original meaning; Table 3 red sentences are examples of meaning/label alteration caused by this synonym replacement process. For example, the original sentence 'you are gay' was a cyberbullying sentence. However, after using M3, a generated sentence 'you are brave' was no more a cyberbullying sentence. Table 4 shows label alteration of 500 samples of cyberbullying and 500 non-cyber bullying samples by manual checking. Results indicate that M1 produced the least and negligible label alteration while the other two methods, M2 and M3, produced an alarming number of label alterations.

Table 2: Size comparison of the expanded and original AskFm datasets as well as the expanded and original FormSpring datasets. O.D. refers to original dataset size, and ExDy refers to expanded dataset using methods 1, 2, and 3.

Dataset Name	Number of Sentences (size)	
	AskFm dataset	FormSpring dataset
Not expanded	10K	12K
ExD1	114k (11 × O.D.)	136k (11 × O.D.)
ExD2	562k (56 × O.D.)	558k (46 × O.D.)
ExD3	1121k (112 × O.D.)	1061k (88 × O.D.)

3.2. Preprocessing

Both cyberbullying datasets and expanded datasets are preprocessed using standard NLP tools, mainly removing unidentified characters, symbols, and tab tokens (e.g., @,‚0-9, #,+, etc.) and converting all characters to lower case. Many abbreviated words and short forms of social

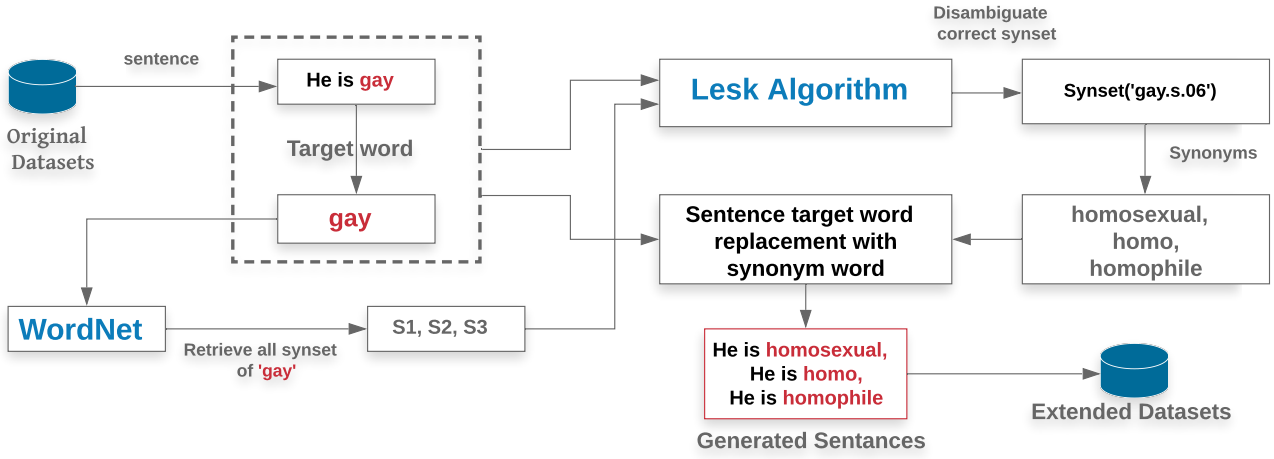

Figure 1: Example of a sentence expansion using proposed Method 1. For a target word, we calculate its corresponding list of synonyms using WordNet. To retrieve only correct synonyms, we insert the synonyms set and the sentence containing the target word to Lesk Algorithm. Once the disambiguation step is done, we generate new sentences by replacing the target word with each of these synonyms to create the expanded dataset.

Table 3: Example of generated sentences from AskFm dataset using M1, M2, and M3. Red sentences represent meaning/label alteration of original sentences during synonym replacement.

Original Sentence	**M1 Generated Sentences**	**M2 Generated Sentences**	**M3 Generated Sentences**
you are gay	you are gay, you are queer, you are homophile	you are gay, you are festal, you are sunny, you are cheery, you are jocund, you are queer, you are homophile	you are gay, you are jocund, you cost gay, you be gay, you are brave, you exist gay, you are jolly you equal gay, you are festive you constitute gay, you are homosexual, you represent gay, you live gay

Table 4: Percentage of label alteration for 500 cyberbullying and 500 non-cyberbullying samples using different synonym replacement methods.

-	**M1 label alteration**	**M2 label alt.**	**M3 label alt.**
Cyberbully Sentence	2%	17%	23%
Non-cyberbully Sent.	1%	11%	13%

network slang are replaced with their original terms (e.g., u = you, em = them, tbh = to be honest, etc.). Besides, stop words, which are generally the most common words in a language, are removed. We have used NLTK's[5] list of English stop words as reference. However, we have modified that lists since some stop words are essential to determine cyberbullying. For example, in some cases, person indicator words (e.g., he, she, his, her, you, yourself, etc.) could be considered stop words; however, those play a crucial role in cyberbullying detection.

3.3. Feature Engineering

A set of features have been employed and evaluated for cyberbullying detection.

TF-IDF. The term frequency (TF) accounts for the absolute frequency of the tokens in the corpus. The TF-IDF considers the rate of each token weighted by its inverse document frequency in the corpus. It reflects how important an individual token is to a document in the database.

N-grams. This assumes a contiguous sequence of N (N>1) tokens instead of a single token of bag-of-word model. Other representations involve N-gram at character level instead of token level. Unlike (word level) TF-IDF, n-gram features allow us to account for the ordering among the tokens. Several combinations of TF-IDF and N-gram features have been tested where n ranges from a lower bound and an upper bound. In our experiment, [2,3] and [3,4]-grams are found to be the n-gram features that best improved the detection after an initial exploration stage. We have used three different combinations of TF-IDF: Word-level, N-Gram word level (for N=2, 3), and N-Gram Character level (for N=3, 4). The Word level TF-IDF feature assigns a score to every term in documents, while the word-level N-gram feature applies TF-IDF scoring to all 2-grams and 3-grams tokens extracted from the whole corpus dataset. The Character level TF-IDF provides a matrix representation of TF-IDF scores of character-level n-grams (n=2, 3) in the corpus. We restricted to 5000 features for each type to avoid the computational burden.

Word Embeddings Features. Word embedding maps each token to a vector of real numbers aiming to quantify and categorize the semantic similarities between linguistic terms based on their distributional properties in a large cor-

[5] https://gist.github.com/sebleier/554280(accessed Dec 30, 2020)

pus using ML or related dimensional reduction techniques. In our case, we have used the pre-trained word embedding; namely, FastText [6]. FastText can be exploited both using its word embedding representation employed as an input to another machine learning architecture and as a text-classifier itself. Therefore, we have run a small experiment on whether it is worth using FastText as an embedding instead of a text-classifier as well. Table 5 shows that FastText as a classifier yielded only 55.4% accuracy and 54.7% F1 scores. In contrast, FastText as word embedding with a CNN classifier yielded 91% accuracy and F1 scores. This outcome motivates the use of FastText as a word embedding for this particular cyberbullying domain.

Table 5: Classifier Accuracy(%) and F1(%) results comparison between FastText as text classifier and FastText as word embedding with CNN for AskFm base dataset.

FastText as classifier	FastText word embedding with CNN
55.4 (Acc.), 54.7 (F1)	91 (Acc.), 91 (F1)

3.4. Classification Architecture

Once our data is preprocessed, we performed the binary cyberbullying classification. Initially, we employed a random split of the original dataset into 70% for training, 10% for validation, and 20% testing. All the results in this study have followed the same test setup. In other words: the original AskFM dataset was first split into 70% train, 10% validation, and 20% testing, and the expansion methods were only applied to the training data while the test data was kept the same for all experiments. A similar procedure was followed for the FormSpring dataset as well. This was very important because if the test data varied from one method to the next, that would be a significant flaw of the methodology. Three types of classifiers were implemented: Convolution Neural Network (CNN), and two baseline algorithms: Logistic Regression and Naive Bayes.

We adopted (Kim, 2014) CNN architecture, where the input layer is represented as a concatenation of the words forming the post (up to 70 words), except that each word is now represented by its FastText embedding representation with a 300 embedding vector. A convolution 1D operation with a kernel size of 3 was used with a max-over-time pooling operation over the feature map with a layer dense 50. Dropout on the penultimate layer with a constraint on l2-norm of the weight vector was used for regularization. Fig. 2 illustrates our CNN architecture.

The details of the implementation are reported on the GitHub page of the project[7]. The various features were examined by each classifier in order to test its accuracy and robustness in classifying hate statements.

3.5. Performance metrics

To demonstrate the performance of our proposal, we calculate the accuracy and F_Measure as follows:

F_Measure: determines the harmonic mean of precision and recall by giving information about the test's accuracy. It is expressed mathematically as:

$$F\_Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (1)$$

Accuracy: measures the percentage of correct predictions relative to the total number of samples. It can be expressed as:

$$Accuracy = \frac{\#TP + \#TN}{\#TP + \#FN + \#TN + \#FP} \quad (2)$$

where #TP, #FN, #TN, #FP correspond to the numbers of true positives, false negatives, true negatives and false positives respectively.

3.6. Transformer Networks Models

BERT – is the Bidirectional Encoder Representations from Transformers: this seminal transformer-based language model applies an attention mechanism that enables learning contextual relations between words in a text sequence (Devlin et al., 2018). Two training strategies that BERT follows:

1. Masked Language Model (MLM): where 15 % of the tokens in a sequence replaced (masked) for which the model learns to predict the original tokens, and
2. Next Sentence Prediction (NSP): in which the model receives two sentences as input and learns whether the second sentence is a successor of the first sentence in their original document context.

3.7. Experiment setup with BERT model

We fine-tuned different transformer models with our AskFm training data using the corresponding test data for validation. The following models were tested: BERT-base and BERT-large (uncased). Each model was fine-tuned for 6 epochs with a learning rate of 5e-6, a maximum sequence length of 128, and batch size 4. After each epoch, the model was evaluated on the validation set.

4. Results

The results of the binary classification of cyberbullying for the original dataset and the expanded datasets are summarized in Tables 6 and 7. Table 6 shows a comparison of classifier accuracy and F1 score for all four types of classifiers with 'AskFm Not Expanded Dataset', 'Expanded Dataset 1', 'Expanded Dataset 2', and 'Expanded Dataset 3' generated by the proposed Method 1, 2, and 3, respectively. We have observed that the CNN classifier outperformed all other classifiers. Therefore, in Table 7, we have only shown the results of the CNN classification using Word-Embedding feature, and used 'AskFm not expanded datasets', 'FormSpring not expanded datasets', cross dataset and All 'expanded datasets' for results comparison. Cross dataset 1 refers to a dataset created using AskFm for training and FormSpring for testing. Similarly, Cross dataset 2 refers to a dataset created using FormSpring for training and AskFm for testing.

[6] https://fasttext.cc/docs/en/crawl-vectors.html (accessed Dec 30, 2020)

[7] https://github.com/saroarjahan/expansion_of_cyberbyllying_datasets/

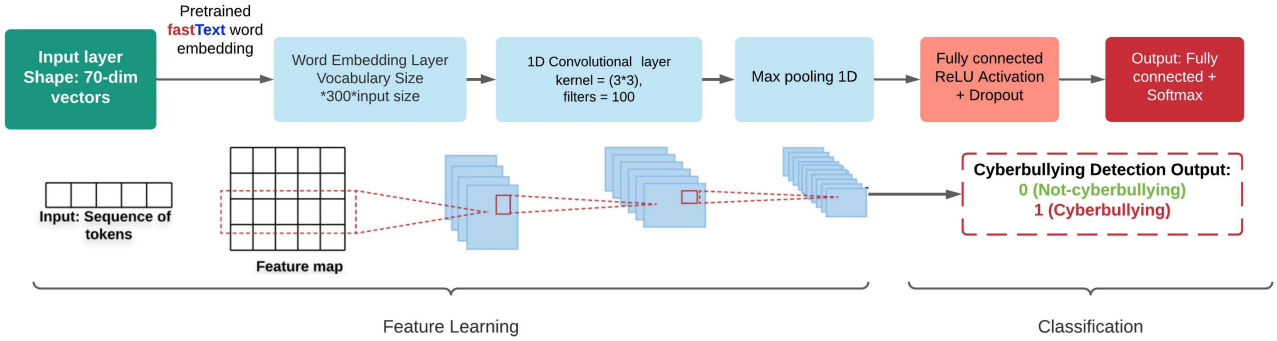

Figure 2: The architecture of our proposed cyberbullying detection using CNN and FastText.

Table 6: Classifier Accuracy (%) and F1 scores (%) for AskFm not expanded dataset, and its underlying expanded datasets using method 1, 2 and 3 respectively.

Classifier and Feature Name	**Not-Ex.**		**Ex. D.1**		**Ex. D.2**		**Ex. D.3**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Naive Bayes + WordLevel TF-IDF	88	82	88.9	82.7	88.5	81.5	88.4	82.3
Naive Bayes + CharLevel TF-IDF	88	83	89.1	84.1	88.7	83.5	88.3	83.4
Logistic regression + WordLevel TF-IDF	90	88	91.2	89.1	90.8	88.6	90.5	88.5
Logistic regression + CharLevel TF-IDF	90.1	89	91.5	91.4	92.2	91.2	91.7	89.6
CNN + Word Embedding	91.2	91	**94.3**	**94.2**	93.6	93.5	93.1	93.1
BERT-large-cased	91.1	89.9	93.7	93.1	92.5	91.5	92.1	91
BERT-large-uncased	91.2	90	93.8	93.2	92.8	91.7	92.3	91.2
BERT-base-cased	91.1	90	93.4	93	92.2	91.3	92.0	90.8
BERT-base-uncased	91.4	91.1	93.9	93.3	92.7	91.7	92.5	91.4

Table 7: Classifier Accuracy (%) and F1 scores (%) of CNN classification using word embedding representation for original and expanded datasets. Cross dataset 1 refers to the performance of the classifier trained using AskFm dataset and tested on FormSpring dataset. Similarly, Cross dataset 2 refers the performance of the classifier trained using FormSpring dataset then tested on AskFm dataset. Also, ExDy, may refers to expanded dataset using methods 1, 2 and 3.

Dataset name	**AskFm**		**FormSpring**		**Cross dataset 1**		**Cross dataset 2**	
	Acc.	**F1 score**	**Acc.**	**F1 score**	**Acc.**	**F1 score**	**Acc.**	**F1 score**
Not expanded	91.1	91	95	94.1	89.2	88.3	76.5	80.7
ExD1, M1	**94.3**	**94.2**	**98.3**	**98.1**	**93.3**	**92.1**	**79.4**	**83.1**
ExD2, M2	93.6	93.5	97.2	97.4	92.5	91.8	78.7	82.6
ExD3, M3	93.1	93.1	95	95	92	91.5	78	81.3

The results highlighted in Tables 6 and 7 indicate the following:

- Among all four types of the classifiers, BERT-based-uncased performed best for non augmented dataset. However, CNN performs best for extended datasets. This indicates that the adopted CNN architecture with the word-sense disambiguation based augmentation strategy worked better.

- For base datasets, when the dataset size is small, all classifiers yield a good accuracy and F1 score. However, when the same classifier was applied to an extended dataset, the CNN showed a clear outperformance compared to the baseline classifiers (NB and LR), which demonstrates its efficiency.

- Among baseline classifiers (NB vs. LR), Logistic Regression models outperform Naive Bayes models in accuracy and F1 scores. However, both LR and NB performed low compared to BERT and CNN. More specifically, CNN 3.1%, Bert best model 2.5%, LR 1.4%, and NB 1.1% performed better than non augmented datasets using M1.

- Among TF-IDF features, 'Character Level TF-IDF' outperformed 'Word Level TF-IDF'.

- In table 7, we observed that all the three proposed methods for data expansion yielded close scores with negligible deviation. However, the proposed M3 has shown .5% less accuracy than M2, and M2 shown .8% compared to M1 for both AskFm and FormSpring datasets. A possible explanation could be that 'M3' and 'M2' covers some meanings that are not relevant to the words as they occur in the text (Tab. 4). Therefore, Methods 1 work significantly better because they have been using word-sense disambiguation tagging capable of targeting more sense-specific synonyms.

- Classification results for extended datasets have been improved way better than classifiers' results for initial AskFm and FormSpring datasets. For CNN, the initial accuracy score has increased from 91% to 94.3% for the AskFm dataset and from 95% to 98.3% for the FormSpring dataset. This improvement is exhibited in all the four other classifiers. This outcome indicates that the semantic meaning expansion using disambiguation and Wordnet worked quite well.

We compared our data augmentation technique to another method called Mixup; both applied on the AskFm dataset. We followed the Mixup implementation provided by (Zhang et al., 2017), where the two random same labeled datasets and their corresponding labels are mixed up to form a new labeled dataset. One can observe from Table 8 that using Mixup augmentation on AskFm dataset yields an accuracy of 91.9% and an F1 score of 91.4%, which is 2.4% lower accuracy compared to our proposed Method 1 for data augmentation.

Table 8: Classifier Accuracy(%) and F1(%) results comparison with Mixup technique, using training and test dataset for AskFm and random word replacement.

Before augmentation	**Augmentation using M1**	**Mixup**	**Random word replacement**
91.1 (Acc.) 91 (F1)	94.3 (Acc.) 94.2 (F1)	91.9 (Acc.) 91.4 (F1)	90.7 (Acc.) 89.2 (F1)

Besides, we conducted other experiments using a different test set (AskFm as test and FromSpring as training) instead of composing the training and test sets from the same corpus. For Cross dataset 1, the accuracy increased from 89.2% to 93.3% when using data augmentation method 1, while the F1 score has improved from 88.3% to 92.1%. In other words, a 4% and a 3% performance improvement in terms of accuracy and F1 score are observed when using our data augmentation approach. Similar performance improvement has been observed for Cross dataset 2 as well. This demonstrates the feasibility and tractability of our developed approach. Furthermore, we used a random word replacement method instead of using the same contextual word or synonym for data expansion. This is to ensure whether our expansion method was not affected by any other irrelevant factor (ex. large datasets work better in ML compared to small datasets). We have observed that this technique yielded an accuracy of 86.7% (Table 8, column 3), which decreases the performance with 5.3% compared to the results of without data augmentation. Therefore, this outcome supports the use of the word replacement technique with contextual meaning and synonyms.

Similar to us, (Zhang et al., 2016) proposed a novel cyberbullying detection with pronunciation based convolutional neural network (PCNN). Since they used fromspring datasets, a light comparison to our results is reported in Table 9. The comparison of these results clearly shows that both CNN and PCNN models by (Zhang et al., 2016) yield a max 96.8% accuracy and 56% F1 scores. However, our CNN models trained on expanded Fromspring datasets using proposed methods 1,2 and 3 yield 1.5% higher Accuracy and 42.1% higher F1 scores.

Table 9: FromSpring datasets results comparison using CNN acrhitecture between (Zhang et al., 2016) and ours expanded FormSpring datasets

Authors name	**Classifier**	**Accuracy**	**F1 score**
Zhang	CNN	96.4	48
Zhang	PCNN	96.8	56
Ours (Expanded FormSpring dataset using Method 1)	CNN	98.3	98.1

5. Conclusion and future work

This paper deals with simple semantic meaning expansion using sense disambiguation for cyberbullying datasets and compares its identification using original feature engineering. The methodology was tested on two different cyberbullying datasets collected from social networks: AskFm and FormSpring, and six artificially expanded datasets. Our technique was also compared to an existing data augmentation approach, Mixup. A convolutional neural network architecture that uses FastText word embedding features and BERT was contrasted to baseline algorithms, constituted of Logistic Regression and Naives' Bayes classifiers. In all cases, BERT and CNN outperformed the baseline classifiers. Furthermore, both CNN and BERT models showed an increase in model performance while using augmented datasets. The testing results demonstrate the feasibility of the extended datasets for semantic meaning expansion, which clearly showed enhanced performance compared to POS tag synonym and general synonym replacement. This experiment answers the fundamental question that targeting all synonym replacements improves the classifier model learning; however, it also largely harms the training data by altering labels. On the other hand, sense-disambiguation 'M1' showed promising results for the lowest label alteration and high performance compared to M2 and M3, which is very promising and would inspire the development of close-meaning augmentation methods.

The superiority of the constructed CNN-BERT model in the overall classification for all datasets is clearly emphasized. Moreover, we believe this work will pave the way for a better-improved identification of bullying intents on social media in a way to guide future training and precaution measures. The disambiguation and the semantic expansion used in this work are not specific only to cyberbullying tasks and, therefore, can be exploited in other text categorization tasks. However, sometimes the same sense of synonyms may alter the meaning for a particular context. We therefore plan to develop more tailored algorithmic schemes that can target suitable synonym class to expand semantic meaning without alteration of context. We also plan to utilize state-of-the-art deep learning word-sense disambiguation approaches to guide this process.

6. Acknowledgements

This work was partially supported by EU Project YougRes on youth polarization & radicalization (ID: 823701) and COST Action NexusLinguarum – "European network for

Web-centered linguistic data science" (CA18209), which are gratefully acknowledged.

7. Bibliographical References

Agrawal, S. and Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *European Conference on Information Retrieval*, pages 141–153. Springer.

Beddiar, D. R., Jahan, M. S., and Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

Bello, I., Zoph, B., Vasudevan, V., and Le, Q. V. (2017). Neural optimizer search with reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 459–468. JMLR. org.

Bu, S.-J. and Cho, S.-B. (2018). A hybrid deep learning system of cnn and lrcn to detect cyberbullying from sns comments. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 561–572. Springer.

Burnap, P. and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Chaplot, D. S. and Salakhutdinov, R. (2018). Knowledge-based word sense disambiguation using topic models. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Chen, Y. (2011). Detecting offensive language in social medias for protection of adolescent online safety.

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

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30.

Ekedahl, J. and Golub, K. (2004). Word sense disambiguation using wordnet and the lesk algorithm. *Projektarbeten 2004*, 17.

Fellbaum, C. (2010). About wordnet. wordnet. princeton university.

Foong, Y. J. and Oussalah, M. (2017). Cyberbullying system detection and analysis. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pages 40–46. IEEE.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Gong, S., Zhang, J., Zhao, P., and Jiang, X. (2014). Tweets and sales. *Available at SSRN*, 2461370.

Guo, H., Mao, Y., and Zhang, R. (2019). Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Gutiérrez, Y., Vázquez, S., and Montoyo, A. (2017). Spreading semantic information by word sense disambiguation. *Knowledge-Based Systems*, 132:47–61.

Jahan, M. S. and Oussalah, M. (2021). A systematic review of hate speech automatic detection using natural language processing. *arXiv preprint arXiv:2106.00742*.

Jungiewicz, M. and Smywiński-Pohl, A. (2019). Towards textual data augmentation for neural networks: synonyms and maximum loss. *Computer Science*, 20.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Kumar, A. and Sachdeva, N. (2020). Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. *Multimedia systems*, pages 1–15.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., and Choo, K.-K. R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. *Concurrency and Computation: Practice and Experience*, page e5627.

Naskar, S. K. and Bandyopadhyay, S. (2007). Word sense disambiguation using extended wordnet. In *2007 International Conference on Computing: Theory and Applications (ICCTA'07)*, pages 446–450. IEEE.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Nockleby, J. T. (2000). Hate speech. *Encyclopedia of the American constitution*, 3(2):1277–1279.

Patchin, J. W. and Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth violence and juvenile justice*, 4(2):148–169.

Reynolds, K., Kontostathis, A., and Edwards, L. (2011). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Rosa, H., Matos, D., Ribeiro, R., Coheur, L., and Carvalho, J. P. (2018). A "deeper" look at detecting cyberbullying in social networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., Coheur, L., Paulino, P., Simão, A. V., and Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior*, 93:333–345.

Sahin, G. and Steedman, M. (2018). Data augmentation via dependency tree morphing for low resource languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5004–5009. ACL Anthology, November.

Salawu, S., He, Y., and Lumsden, J. (2017). Approaches to automated detection of cyberbullying: A survey. *IEEE Transactions on Affective Computing*.

Sun, X. and He, J. (2020). A novel approach to generate a large scale of supervised data for short text sentiment analysis. *Multimedia Tools and Applications*, 79(9):5439–5459.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., and Hoste, V. (2015). Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 672–680.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26. Association for Computational Linguistics.

Zhang, Z. and Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.

Zhang, X., Tong, J., Vishwamitra, N., Whittaker, E., Mazer, J. P., Kowalski, R., Hu, H., Luo, F., Macbeth, J., and Dillon, E. (2016). Cyberbullying detection with a pronunciation based convolutional neural network. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 740–745. IEEE.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.