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@oulu.fi

Abstract

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. 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.

Objectives

In overall, the main contributions of this work are as follows:

(i) Three different synonym-based augmentation schemes have been put forward.

(ii) We compared our proposed data augmentation approach against the state-of-art Mixup, a newly proposed data augmentation method

(iii) We developed a new python library for data augmentation, which is the end product of our experiment, and released under an open-sourced

(iv) We released a new cyberbullying dataset and considered the subtle differences between common hate speech and cyberbullying during annotation.

METHODOLOGY

Our experiment methodology includes a three-fold process.

1. A newly collected dataset from AskFm social network website and the publicly FormSpring dataset were introduced.
2. Next, data augmentation is performed using the Wordnet-based sense disambiguation technique and Lesk-algorithm.
3. Then, classification approach involving BERT, CNN, NB, LR, and FastText models. The results are contrasted and the data augmentation process has been duly evaluated for both AskFm and FormSpring datasets.

AskFm Dataset Creation

The first original dataset that we have used in this paper is collected from Ask.Fm website. We have crawled each of the profiles using Python web crawler library. 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, We have manually labeled the resulting 10k AskFm dataset. Labeling involves identifying whether each sentence contains cyberbullying or not.

Data Augmentation Methods

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 was employed.

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.

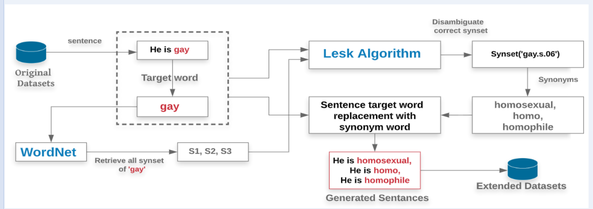

Figure1: Data augmentation Method 1.

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

Classifier Architecture

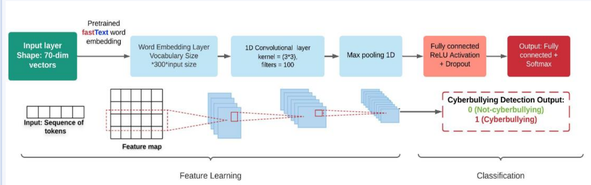

Figure 2: The architecture of our proposed cyberbullying detection using CNN and FastText.

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 types of classifiers were implemented for all experiments.

Convolution Neural Network (CNN), and two baseline algorithms: Logistic Regression and Naive Bayes. We adopted CNN architecture by kim [1], 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.

RESULTS

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

Conclusion

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.

References

[1] Kim, Y. (2014). Convolutional neural networks for sentence classification.

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.