**Automatic Detection of Cyberbullying on Social Networks**

**based on Bullying Features**

Rui Zhao

rzhao001@e.ntu.edu.sg

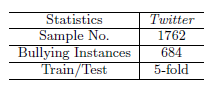
Anna Zhou

annazhou89@gmail.com

Kezhi Mao

In this paper, we propose a representation learning framework specific to cyberbullying detection. Based on word embeddings, we expand a list of pre-defined insulting words and assign different weights to obtain bullying features, which are then concatenated with Bag-of-Words and latent semantic features to form the final representation before feeding them into a linear SVM classifier.

**Dataset**

The public bullying traces dataset is adopted here3, which consists of tweets that are messages sent on Twitter [17]. The Twitter dataset is composed of tweets crawled by the public Twitter stream API. Each tweet contains at least one of the following keywords: bully, bullied, bullying. Re- tweets are removed by excluding tweets containing the acronym `'RT`'. Finally, 1762 tweets are sampled uniformly from the whole tweets collections on August 6, 2011 and manually labeled.

**Preprocesses**

To preprocess these tweets, a twitter-specialized tokenizer5 is applied without any stemming or stopwords removal operations. In addition, some special characters including user mentions, URLS and so on are replaced by predefined characters, respectively. The statistics of this dataset can be found in Table 2. Since the dataset does not have explicit train/test split, 5-fold cross validation (CV) is applied, where four-folds are used for training and the remaining one is used for testing. The mean results will be reported.

**Experimental Setup**

The following methods will be compared:

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

For BoW, the most frequent 2000 terms including unigram and bigram are used as features.

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

\* LSA: Latent Semantic Analysis [10].

\* LDA: Latent Dirchilet Allocation [1]. Our implementation of LDA is based on Gensim6.

For LSA and LDA, the number of latent topics are both set to 100. To implement LDA, we set hyperparameter \_for document topic multinomial and hyperparameter \_ for word topic multinomial to 1 and 0:01, respectively.

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

For our proposed method, BoW and latent semantic features are both the same as those in two above compared models. For bullying features, the number of insulting seeds is 20. The number of similar words compared to each in-

sulting seed h is set to 50. After feature expanding based on word embeddings, the length of bullying features is 64 finally. Therefore, the dimension of features learned by our EBoW is the sum of the lengths of the three kinds of features as: 2000 + 100 + 641 = 2741.

**Feature extraction**

Bag-of-Words Features

To extract Bag-of-Words features, a vocabulary including unigram and bigram is constructed firstly and the terms whose document frequencies are less than 2 are all ignored. Different term weighting schemes including tf-idf and binary ones can be applied here [3, 16]. In this paper, we adopt the

tf-idf weighting scheme.

Latent Semantic Features

Here, latent semantic features refer to the features extracted by Latent Semantic Analysis (LSA) [10]. In principle, LSA applies Singular Value Decomposition (SVD) on the term-document matrix that each column is the above

Bag-of-Words features. Then, the derived latent space is spanned by dominant eigenvectors corresponding to large eigenvalues. Each new feature is a linear combination of all original features.

Bullying Features

Insulting words can be pre-defined and extracted based on our prior knowledge and other public linguistic resources, which are named insulting seeds in this paper. Then, we extend these insulting words automatically based on word embeddings

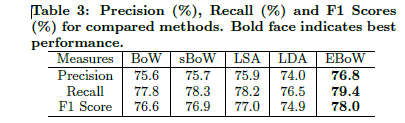
**Classifier**

We then apply linear SVM [2] in the new feature space generated by the above mentioned approaches. In linear SVM, the hyper-parameter: regularization term C, is searched over a range as f2􀀀2; 2􀀀1; 20; 21; 22; 23; 24; 25g via a \_ve-fold cross-validation conducted on training data.

**Experimental Results**

As a result, our approach is able to gain a significant performance improvement compared to sBoW over

All three evaluation measures.

****

The reduced dimension is a key parameter to determine the quality of learned feature space. Here, we \_x the dimension of latent space to 100. A deliberate searching for this parameter may boost the performances of

LSA and LDA

**Parameters Sensitivity**

In our proposed EBoW model, we select the top-h similar words for each pre-difined insulting word. Parameter h is used as the threshold to expand bullying features. An increased h will mean more terms are considered as bullying

features so that the dimension of bullying features increases.

In this section, we investigate the inuence of the parameter on our model performance for cyberbullying detection.

Parameter h is chosen from a prede\_ned set: [0, 50, 100, 150, 200, 250]. When h = 0, bullying features only consist of our pre-defined insulting words. According to the total 6 settings of h, we feed our corresponding learned EBoW features into the classifier. The other experimental settings are kept unchanged. Then, precision, recall and F1 scores for these different settings are calculated and reported in Figure

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