Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7040–7048
Marseille, 20-25 June 2022

Deep One-Class Hate Speech Detection Model

Saugata Bose, Guoxin Su
University of Wollongong
Northfields Ave, Wollongong, Australia
sb632@uowmail.edu.au, guoxin@uow.edu.au

Abstract

Hate speech detection for social media posts is considered as a binary classification problem in existing approaches, largely neglecting distinct attributes of hate speeches from other sentimental types such as "aggressive" and "racist". As these sentimental types constitute a significant major portion of data, the classification performance is compromised. Moreover, those classifiers often do not generalize well across different datasets due to a relatively small number of hate-class samples. In this paper, we adopt a one-class perspective for hate speech detection, where the detection classifier is trained with hate-class samples only. Our model employs a BERT-BiLSTM module for feature extraction and a one-class SVM for classification. A comprehensive evaluation with four benchmarking datasets demonstrates the better performance of our model than existing approaches, as well as the advantage of training our model with a combination of the four datasets.

Keywords: BERT, BiLSTM, One-Class SVM, outlier detection, transfer learning, hate class.

1. Introduction

Hate speech detection in social media posts differs from information retrieval from traditional documents due to length constraint. Moreover, the challenge mounts up due to insufficient publicly available hate speech datasets (MacAvaney et al., 2019), the absence of benchmark datasets (Swamy et al., 2019), and the fuzzy boundary between hate speech and cyberbullying, abusive language, discrimination, profanity, toxicity, flaming, extremism and radicalization concepts (Fortuna and Nunes, 2018; Poletto et al., 2020).

Early works modelled the hate speech detection as a binary classification problem. Since non-hate posts contain a variety of sentiment types which do not share the same characteristics, forcing those samples into one class, as opposed to the hate class, leads to low performance. In view of this, we consider this problem as a one-class problem, where we determine which instances stand out as being dissimilar to the hate class. Those instances are treated as outliers and can be identified effectively by a one-class model such as One-class Support Vector Machine (OC-SVM) (Schölkopf, 2001).

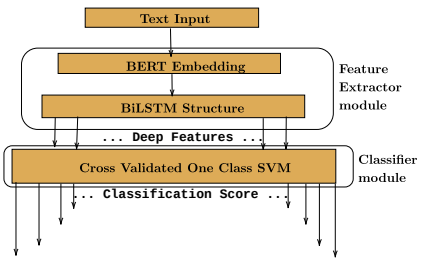

Figure 1: Deep One-Class Hate Speech Detection Model.

In this paper, we propose a novel hate speech detection model (see Figure 1) which includes a deep learning network for feature extraction and the one-class SVM for classification. The deep learning network that we use constitutes the Transformer-based encoder architecture BERT and a Bi-directional Long Short-Term Memory (BiLSTM) model. We evaluate our model with four publicly available datasets: Davidson (Davidson et al., 2017), SemEval-2019 (Task-5, Subtask-A) (Basile et al., 2019), HASOC-2019 (Subtask B) (Mandl et al., 2019) and Stormfront (de Gibert et al., 2018).[1] Moreover, we combine the four datasets to experiment with the generalizability of our model.

We show that our model achieves better performance in most scenarios than the state-of-the-art methods for monolingual datasets.

Our main contributions are summarized as follows:

- **One-class detection model combined with deep learning approach outperforms a binary-class detection model.** We cast a new viewpoint on the weakness of the binary classification approach in hate speech detection and demonstrate our claim by a performance comparison between one-class classification and binary classification.
- **Deep One-Class Hate Speech Detection Model outperforms the baseline models.** We carry out extensive experiments which convincingly demonstrate that our model outperforms other state-of-the-art deep learning approaches to hate speech detection.
- **Deep One-Class Hate Speech Detection Model is generalized.** Our experiments show that the proposed model achieves significantly improved performance after being trained with a combination of four datasets.

[1] We only consider English posts in those datasets.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 describes our model. Section 4 elaborates the experimental setup. Section 5 analyzes experiment results. Section 6 discusses our model. Finally, Section 7 concludes our work and discusses future directions.

2. Related Work

Although the first research article on hate speech detection from public opinion dates back to early 2010s (Warner and Hirschberg, 2012), this research is still at an early with limited and specialized improvement over the years (Arango et al., 2019). Yet, it is growing popular in the NLP community (Poletto et al., 2020). When the question comes to detection, most researchers prefer to implement supervised machine learning based approaches dominated by linear classifiers (Davidson et al., 2017; Malmasi and Zampieri, 2017) or deep learning classifiers (Gambäck and Sikdar, 2017; Badjatiya et al., 2017; Fortuna and Nunes, 2018; Zhang et al., 2018; Founta et al., 2019) or combination of both (Nobata et al., 2016; Paschalides et al., 2020).
Transformer-based Pre-trained Language Models are among the deep learning models which are right now claiming state-of-the-art performance (Minaee et al., 2021). The variants of BERT (Devlin et al., 2019) are one those models which are current trend in hate speech detection. (Mozafari et al., 2019; D'Sa et al., 2020; Fortuna et al., 2021) are few of the current BERT based experiments on hate speech detection. (Mishra and Mishra, 2019) has claimed a BERT based winning model in HASOC-2019 competition. (Mozafari et al., 2019; MacAvaney et al., 2019; Alonso et al., 2020; Ranasinghe et al., 2021) have reported state-of-the-art performance after experimenting BERT variants with Davidson, Stormfront and HASOC-2019 (Sub-task B) and SemEval-2019 respectively. Apart from experimenting with BERT, we observe a similarity in those models. All have considered hate speech detection as a binary problem- texts which "belong to hate" category and texts which do "not belong to hate" category.

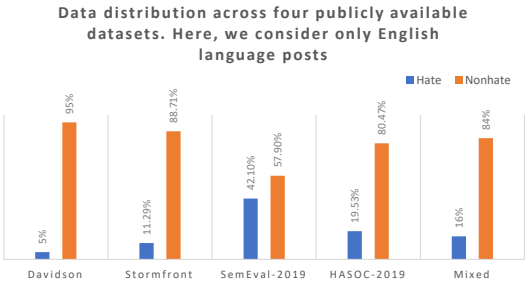

Figure 2: Data Distribution.

If we closely look at Figure 2, we notice the fragility of such strategy. The texts which do not belong to "hate class" are the majority of those datasets. Here, not belongs to "hate class" means anything other than than "hate class" as "hate speech" are very distinctive from "offensive", "aggressive" or "abusive" classes (Yin and Zubiaga, 2021). Early models have been dominated by these non hate instances during training. Because the training set does not resemble the 'true' distribution, the generalization performance is poor as well. Hate speech detection model expects the classifier must know only one class, which is the hate class and its features. The other instances will be treated as anomalies or outliers as their features deviates significantly from other observations (Chalapathy et al., 2019).

This study considers *hate speech classification as a one class classification problem* where the classifier will be trained with only hate class features' and will be able to detect anomalies during validation and test phase while the prediction differs from actual state. We introduce a novel point of view to resolve the detection task. As deep neural network classifiers cannot be trained with only one class, one-class SVM (OC-SVM) is widely used in this regard where a hyperplane separates the positive class (Schölkopf, 2001), in this study which is hate class. OC-SVM will be described in Section 3. In recent times researchers have utilized combination of deep learning and one class classifier in tasks as diverse as visual, speech anomaly detection (Chalapathy et al., 2019) where deep features were extracted using autoencoder, pre-trained transfer learning models and then they fed the features one-class SVM classifier (Pan and Yang, 2010; Andrews et al., 2016; Sun et al., 2017). However, in our proposed *Deep One-Class Hate Speech Detection model*, we obtain deep contextual features from a pre-trained BERT-BiLSTM module and fed those learned features into a one-class SVM (OC-SVM) classifier to separate all the positive data points from the origin. Transferring knowledge from BERT to BiLSTM is not new. (Tang et al., 2019) uses a metaphor to explain an architecture where, BERT the large model serves as a *teacher* and BiLSTM, the small model learns to mimic the teacher as a *student* where a student reproduces the behavior of the teacher as accurately as possible and with fewer parameters. This architecture has shown improved accuracy and proven "more expressive for natural language tasks" (Tang et al., 2019).

Generalisability of a model has also become a concern. (Gröndahl et al., 2018) reckons a model performs well only when trained and tested on the same dataset. The model experiences improvement when more training data were added (Fortuna et al., 2018). These studies revealed that a model trained with a large number of hate samples could be competitive with the state of-the-art models with the additional benefit of allowing prediction on cross datasets as well.

We draw inspiration from these existing studies and build a deep neural network joint by one-class SVM for hate speech detection.

3. Deep One-Class Hate Speech Detection Model

Our deep one-class model for hate speech detection (Figure 1) comprises two modules, namely, a feature extractor module (Figure 3) and a classifier module (Figure 4).

3.1. Feature Extractor by Transfer Learning

We use the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) for extracting deep linguistic features, which provide different representations for the same words with meanings. Then, we feed the output of BERT to a low dimensional sequence model BiLSTM to capture the long-term dependency of words. For classification, we just use a dense layer with the cross-entropy loss. Figure 3 presents the architecture of the two module in detail.

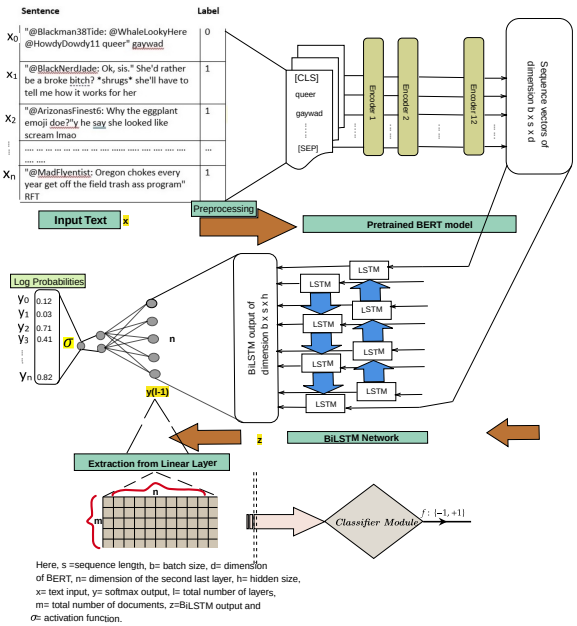

Figure 3: Feature Extractor Module maps $x \in \mathbb{R}^d$ into $y \in \mathbb{R}^{d_1}$, which is extracted from the second last layer.

Given a training dataset $D = \{(x_i, c_i) \mid i = 1, \ldots, n\}$ where $x_i \in \mathbb{R}^{d_0}$ is a vector representation of a document (i.e., each word is represented by a unique integer) and $c_i \in \{0, 1\}$ is the label (1 for the hate class and 0 otherwise) for each i.

The pre-trained BERT model projects each x_i into a $y_i \in \mathbb{R}^{d_1}$ where $d_1 =$ BERT dimension. Later, a sequence processing BiLSTM neural network is used to learn the "high-level" features $y_{(l-1)}$ where $y_{(l)}$ is defined in Eq. (1).

$$y_{(l)} = \text{softmax}(y_{(l-1)}) = \frac{e^{y_{(l-1)_i}}}{\sum_{i=1}^{2} e^{y_{(l-1)_i}}} \quad (1)$$

Here $y_{(l)}$ refers the output from last layer (i.e., the softmax layer). Intuitively, for each i, $y_{(l)_i} \in (0, 1)$ is the predicted probability that the input sample belongs to the i-th class. The cross entropy loss is calculated by Eq. (2).

$$loss = -\big(c \log(y_{(l)_i}) + (1 - c) \log(y_{(l)_i})\big) \quad (2)$$

where $c \in \{0, 1\}$ is the target label. After the model becomes trained, the learned feature matrix $Z = \{z_{i,j}\} \in \mathbb{R}^{m \times n}$, m are number of samples, n represents embedding dimension, is obtained from the second last layer of the module as follows:

$$y_{(l-1)} = w^T z_{i,j} \quad (3)$$

Here $z_{i,j}$ is the hidden state of the BiLSTM network at timestep i for the j-th sample, and w represents the final learned weights.

3.2. Classifier module: One-Class SVM

The unsupervised classifier module is summarized in Figure 4. Unlike binary classifiers, one-class classifier is trained to differentiate between positive class data and other instances (Fernández et al., 2018) where the latter are neither present nor properly sampled. In this study, we argue that unlike any variant of non "hate class" instances, hate speeches are properly sampled and carrying specific meanings. So, the module will be trained with the positive class data to detect outlier data and it will be done by constructing a smooth boundary around the majority of probability mass of data.

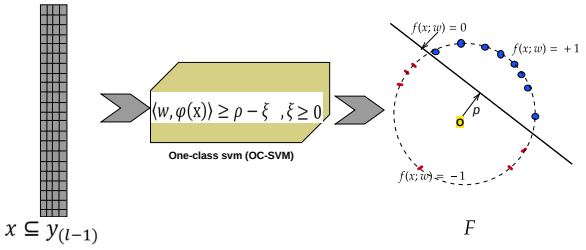

Figure 4: Classifier Module maps datapoints $x \in \mathbb{R}^d$ in *reproducing kernel hilbert space*, F including a hyperplane separation between red (outliers) datapoints and blue (positive) datapoints from the origin O.

The module utilizes one class svm (OC-SVM) classifier, proposed by (Schölkopf, 2001) which defines the hyperplane to discriminate the positive class from the other instances with maximum margin. Through this module we intent to learn a function, f which we can apply to any feature vector $x \in \mathbb{R}^d$. $f(x)$ results a probability that the example x is positive.

$$f : \mathbb{R}^d \to \{-1, +1\} \quad (4)$$

Eq. (4) returns +1 in a "small" region (capturing the positive data points) and –1 elsewhere.

More specifically, given a training data with positive class $x \subseteq y_{(l-1)}$, the optimization objective can be

written as:

$$\min_{w,\rho,\xi} \quad \frac{1}{2} \parallel w \parallel^2 -\rho + \frac{1}{vn}\sum_{i=1}^{n}\xi_i \qquad (5)$$
$$\text{s.t.} \langle w, \varphi(x_i)\rangle \geq \rho - \xi_i, \xi_i \geq 0, \forall i$$

In Eq. (5) and in Figure 4 each of non-negative $\xi = \{\xi_1, \xi_2, ..., \xi_n\}$ slack variables are penalized in the objective function, ρ points the distance from the origin to hyperplane w, φ is is a mapping function that maps x to a a RKHS map function F, $f(x) = w^T\varphi(x_n)$ is a hyper plane decision function to separate as many as possible of the mapped vectors $\varphi(x_i), i : 1, 2, ..., n$ from the origin O and $v \in (0, 1)$ is a trade-off parameter.

4. Experimental Setup

In this section we present the experimental setup for *Deep One-Class Hate Speech Detection Model* to address our contributions. As the primary concern is to learn the features of "hate class", we consider presence of other variants of non "hate class" instances as a single class. The dataset is divided as 80:20:20 ratio and the reported performances are calculated on test dataset.

Convert to two classes (hate one and non-hate one)	Punctuation, Extra space, Irrelevant characters, Unicode, Stop words, Emoticon, Numbers, URL removal
Tokenization	Infrequent tokens removal (empty tokens or the tokens having one character only.)

Table 1: Pre-processing Techniques.

4.1. Experimental Setup: Feature Extractor module

After applying a range of pre-processing techniques listed in Table 1, input text is fed to a small pre-trained BERT model *bert-base-uncased* that contains an encoder with 12 layers (transformer blocks), 12 self-attention heads, and 110 million parameters. During fine tuning, we notice that the module optimizes for 30 epochs with 32 batch size. The following hyper-parameters were set during training: Adam optimizer with learning rate 0.00002 and combination of Log-Softmax and negative log-likelihood loss as a cross entropy gives best performance.

4.2. Experimental Setup: Classifier module

OC-SVM classifier will be trained with only the hate speech feature sets, which is $x \subseteq y_{(l-1)}$. Here, x are only the positive sets. The classifier focuses on outlier detection. As v-property (Schölkopf, 2001), $v \in (0, 1)$ allows one to incorporate a prior belief about the fraction of outliers present in the training data into the model (Ruff et al., 2018), we experiment the module with a range of outliers and we notice different optimized v values for different datasets which have been reported in Table 2. Gaussian kernel was employed to develop the hyperplane. The model was verified through a 10 fold cross validation. We conservatively relabel any documents that are having negative scores. We iterate this process until it stabilizes.

4.3. Datasets

The proposed model has been applied on four publicly available datasets. Although they represent different domains, they carry "hate class" labelled speeches along with other classes. The dataset distribution was presented at Figure 2.

- Davidson dataset: This is one of the most popular English datasets prepared by (Davidson et al., 2017) comprising 24,802 tweets.
- StormFront dataset: This dataset contains texts from Stormfront (de Gibert et al., 2018), a White Supremacy Forum. There are 10,944 posts in this dataset.
- SemEval-2019 dataset (Sub-task A): 10k tweets have been collected for SemEval-2019 (Basile et al., 2019) to solve Task 5 which is about the detection of hate speech against immigrants and women in Spanish and English tweets.
- HASOC-2019 dataset (Sub-task B): The dataset (Mandl et al., 2019) comprises of 5852 facebook and twitter posts; were prepared to solve Sub Task-B which is about to identify hate speeches.
- Mixed Dataset: We customized a dataset by combining Davidson, Stormfront, SemEval-2019 and HASOC-2019 datasets to experiment with cross-domain and cross dataset generalisability of our model. The dataset consists of around 50k English only posts where 16% posts are hate speech.

In this study, we experiment with monolingual data-especially posts in English .

4.4. Methods Compared

We compare our proposed model with the following baseline methods:

- LSTM, BiLSTM: The LSTM and BiLSTM models were created with 256 hidden units and 768 input size. Here, *bert-base-uncased* was used for word embedding.
- CNN: As per formulation in (Zhang et al., 2018). The embeddings' extracted as the sequence output of *bert-base-uncased* were fed to the CNN neural network.

- BERT: A simple BERT-base classifier consisting of a dense layer with 768 feature inputs.

During experiment, CNN and LSTM replaces the BiLSTM in Figure 3. The pooled output from BERT was used in BERT method.

4.5. Performance Metric

In this study, we evaluate the solutions by F1 score of hate class on a balanced dataset which has been summarized in Eq. (6), Eq. (7), Eq. (8).

$$Precision_{(hate)} = \frac{TP_{(hate)}}{TP_{(hate)} + FP_{(hate)}} \quad (6)$$

$$Recall_{(hate)} = \frac{TP_{(hate)}}{TP_{(hate)} + FN_{(hate)}} \quad (7)$$

$$F1_{(hate)} = 2 \times \frac{Precision_{(hate)} \times Recall_{(hate)}}{Precision_{(hate)} + Recall_{(hate)}} \quad (8)$$

Here, TP, FP and FN represent true positive, false positive and false negative respectively. As there are no previous studies of one class classification on hate speech detection, it is not feasible to compare the performance with past models. But our results will give a notion of the credibility of the proposed model.

5. Results

We have conducted a number of experiments to figure out how does one class classifier behave with the hate class features and non hate class features. Among 10 experiments on each dataset, 5 were trained upon an input set containing 95% of hate class and remaining are the non-hate ones. Another 5 experiments were trained upon 100% of the hate class.

As no one has ever tried to implement one-class classifier upon hate speech classification, this is difficult to compare the scores. Analysing the performance of the model demands how well the model detects "hate class". Based on this criterion, we notice that, the best F1 score on Davidson dataset is 0.85 with 2% outliers. With 3% outliers, HASOC-2019 and Mixed dataset have top F1 scores, 0.60 and 0.82 respectively. Similarly with 4% outlier, the model produces best f1 scores for Stormfront (0.88) and SemEval-2019 (0.84). However the model does not produce a significant score for HASOC-2019 dataset comparing with F1 scores of other datasets. Contrasting with other datasets, this one represents a mixed domain (i.e. combining facebook and twitter posts). We notice that HASOC-2019 dataset has a substantive number of hashtags (see Figure 5) which deserves additional attention during feature extraction round. We believe the cause of the poor performance in HASOC-2019 dataset is a weak correlation with the semantics of the speech. Future research will address this issue by proposing a feature based on hate speech lexicon.

Figure 5: Word cloud: HASOC-2019 dataset after preprocessing. It shows hashtags (i.e. trumpisatraitor, shameonicc and so on) are appearing most frequently comparing *rapist, douchebag*.

The experimental results in Table 2 have shown that the incorporation of negative samples during training has significantly improved the performance-the model trained with 95% of hate class outperforms those models trained with 100% hate speech features.

6. Discussion

At the beginning of this study, we argue that "hate class" labeled speeches have distinctive features compared to non "hate class" instances. Our approach differs from the past studies where researchers' have considered texts belong to "hate class" and not belong to "hate class" are two classes-which ends up in a binary detection problem. In this study, we argue that the texts which do not belong to "hate class" are undefined as non "hate class" instances do not share any similarities. That makes the hate speech detection a one class problem- a fundamental turnaround from the early models.

In favor of our argument we present a quantitative comparison between binary classification and one class classification in Table 3. The F1 scores are calculated from Eq. (8). Further, we train one-class classifier based model with a mixed dataset and test the model with Davidson, Stormfront, SemEval-2019 and HASOC-2019 datasets. The performance was recorded in Table 4. The bold number represents the best performance in both tables. The italic value with underline represents the second best.

6.1. Hate speech detection: a one-class problem

Table 2 shows that one class classification outperforms binary classifier's F1 value. BiLSTM, Bert-base, LSTM and CNN methods have been trained and tested with Davidson, Stormfront, SemEval-2019, HASOC-

Experiments	Davidson	Stormfront	SemEval'2019	HASOC'2019	Mixed
95% 1 class, 5% 0 class, 5% outlier	0.83	0.87	0.82	0.58	0.81
95% 1 class, 5% 0 class, 4% outlier	0.83	**0.88**	**0.84**	0.58	0.81
95% 1 class, 5% 0 class, 3% outlier	0.83	0.86	0.82	**0.60**	**0.82**
95% 1 class, 5% 0 class, 2% outlier	**0.85**	0.86	0.83	0.55	0.80
95% 1 class, 5% 0 class, 1% outlier	0.82	0.82	0.81	0.54	0.80
100% 1 class, 0% 0 class, 5% outlier	0.67	0.70	0.68	0.44	0.73
100% 1 class, 0% 0 class, 4% outlier	0.69	0.70	0.67	0.44	0.73
100% 1 class, 0% 0 class, 3% outlier	0.69	0.72	0.69	0.45	0.75
100% 1 class, 0% 0 class, 2% outlier	0.70	0.71	0.68	0.45	0.75
100% 1 class, 0% 0 class, 1% outlier	0.70	0.71	0.68	0.46	0.75

Table 2: F1 scores of hate class using *Deep One-Class Hate Speech Detection Model*. 1 represents "hate class" and 0 is for non hate instances. Row represents the 10 experiments and columns represent the dataset. Experiments column shows distribution of "hate class" and outlier in a dataset during one-class training. The best results are highlighted in bold.

2019 and Mixed datasets. All methods use *bert-base-uncased* as word embedding. Each method follows two solving approaches- one is traditional binary classification problem (F1 scores are listed under *2class* column) and other is considering hate speech detection a one class problem (F1 scores are listed under *1class* column). We also record the optimized outlier value for which we achieve best F1 score for one class classification.

We notice that one class classification approach improves the F1 score significantly compared to the binary classification approach across datasets. More specifically, BiLSTM with one class approach improves f1 score by as much as 3% for Stormfront dataset, 2% for Mixed dataset, SemEval-2019 dataset and for Davidson dataset whereas Bert-base classifier with one class approach outperforms others by 4% for HASOC dataset. This performance arguably proves that one-class classifier is able to separate large number of positive class samples from the origin and that shows better performance comparing binary classification.

We observe that all but CNN shows improvement in one class classification approach. CNN based models experience a significant performance drop for most of the datasets. Although we are not investigating the reasons in this paper, we believe a rigorous experiment with hyper-parameters might improve the scores.

While the relatively small improvement in some datasets could be due to the short text nature of tweets or for limited hate speech data, the consistent gain in F1 score suggests that hate speech detection should be considered as a one class classification task rather than a binary classification task.

6.2. Proposed model vs. other models

Table 3 shows that one class classification outperforms binary classification, especially the BiLSTM based one class model- which is our proposed *Deep One-Class Hate Speech Detection Model*. The model has shown improved F1 scores for Davidson (0.85), Stormfront (0.88), SemEval-2019 (0.84) and Mixed (0.82) datasets respectively though produces a competitive score (0.60) for the HASOC-2019 dataset. This performance arguably supports the superiority of the "Deep One-Class Hate Speech Detection Model" over other models.

A closer look at the Table 3 also reveals that BiLSTM model with *bert-base-uncased* embedding produces the second highest score for the Stormfront (0.85), SemEval (0.82) and for HASOC-19 dataset (0.62). These scores have strengthen our claim that proposed feature extractor module (see Figure 3) is able to fetch deep features. Adding the classifier module (see Figure 4) produces competitive scores.

Table 2 shows that unlike other models, CNN fails to produce competitive results. This is also noticeable that the LSTM model with *bert-base-uncased* encoding provides second highest score (0.83) for the Davidson dataset and Bert-base classifier outperforms other models in HASOC-19 dataset (0.66). These scores show us the suitability of the pre-trained BERT model in hate speech detection. An ambitious approach could lead to a significant improvement if BERT model be trained with hate speech dataset.

6.3. Generalized performance

In the round of experiments, the best model was used to test generalisability across the other datasets. The evaluated results were presented in Table 4 which shows that the proposed model improves F1 score after being trained with the mixed dataset. Comparing to the Table 3, the improvement is significant for Davidson dataset (4%) and for SemEval-2019 (3%) dataset. The model shows a slight improvement for Stormfront dataset (1%). However, the performance drops significantly for the HASOC-2019 dataset. Apart from HASOC-2019 result, these performances are in line with a similar conclusion by (Fortuna et al., 2018).

If we closely look at these scores, we notice how

Model / Dataset	BiLSTM			BERT-base			LSTM			CNN		
	2class	1class		2class	1class		2class	1class		2class	1class	
Storm front	*0.85*	Outlier =0.04	**0.88**	0.8	Outlier =0.01	0.8	0.84	Outlier =0.04	0.8	0.83	Outlier =0.01	0.74
Sem eval	*0.82*	Outlier =0.04	**0.84**	0.74	Outlier =0.03	0.77	0.81	Outlier =0.03	0.81	0.81	Outlier =0.01	0.71
David son	0.82	Outlier =0.02	**0.85**	0.75	Outlier =0.01	0.8	*0.83*	Outlier =0.02	0.81	0.72	Outlier =0.01	0.71
HAS OC	*0.62*	Outlier =0.03	0.6	0.6	Outlier =0.02	**0.66**	0.5	Outlier =0.04	0.58	0.59	Outlier =0.03	0.52
Mixed	*0.80*	Outlier =0.03	**0.82**	0.65	Outlier =0.01	0.71	0.80	Outlier =0.01	0.79	0.65	Outlier =0.05	0.65

Table 3: F1 score of hate class for different methods on different dataset (using the *bert-base-uncased* word embedding). They have been trained and tested with the same dataset. The best results are highlighted in bold. The second best scores are italicized and underlined. Combination of BiLSTM-1class represents *Deep One-Class Hate Speech Detection Model*.

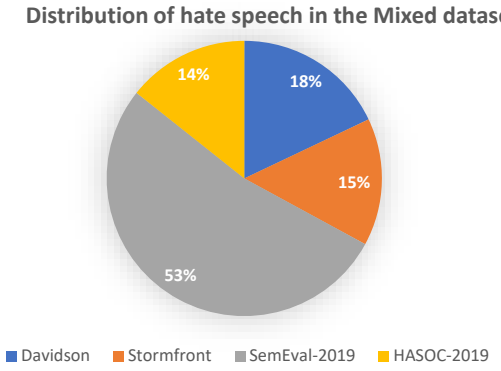

Figure 6: Distribution of positive class data in the mixed dataset.

well the model performs on Twitter based datasets i.e. Davidson and SemEval-2019 dataset. An interesting observation is that both of these datasets are twitter based and a large percentage of positive class samples from SemEval-2019 are in the Mixed dataset (see Figure 6). The same observation might explain the performance dip for HASOC-2019 and Stormfront. Both represents dissimilar domains and both share comparatively small percentage of "hate class" samples in Mixed dataset.

From this experiment, we can conclude that *Deep One-Class Hate Speech Detection Model* is able to show generalisability if trained with a large number of positive samples from the same domain.

7. Conclusion

The paper has proposed a novel framework for hate speech detection. Through comprehensive experiments, we found that if the dataset is balanced, and if the hate class detection becomes the priority, then a one-class classifier will be a best option. We experimented with several state-of-the-art methods and with publicly available datasets. Our results demonstrated that our *Deep One-Class Hate Speech Detection Model* offered the best detection and generalization results.

In terms of future work, we will integrate the classifier module into the neural network architecture which can enable us to influence representational learning in the hidden layers. Furthermore, we will evaluate our ensemble architecture on multidomain-multilingual settings.

8. References

Alonso, P., Saini, R., and Kovács, G., (2020). *Hate Speech Detection using Transformer Ensembles on the HASOC dataset*. Springer International Publishing.

Andrews, J. T. A., Tanay, T., Morton, E. J., and Griffin†, L. D. (2016). Transfer representation-learning for anomaly detection. In *Proceedings of the 33 rd International Conference on Machine Learning*.

Arango, A., Pérez, J., and Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings.Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation*.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017). Deep learning for hate speech detection in tweets.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., and Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings.SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter*. Association for Computational Linguistics.

Chalapathy, R., Menon, A. K., and Chawla, S. (2019). Anomaly detection using one-class neural networks. *arxiv*.

Tested With / Trained with	F1 score			
	Davidson	Stormfront	SemEval'2019	HASOC'2019
Mixed	**0.89**	**0.90**	**0.87**	**0.67**
Davidson	0.85	–	–	–
Stormfront	–	0.89	–	–
SemEval'2019	–	–	0.84	–
HASOC'2019	–	–	–	0.60

Table 4: Cross dataset test results. Rows show the dataset used to train the model and columns represent the dataset used for testing. The best results are highlighted in bold.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings.Automated Hate Speech Detection and the Problem of Offensive Language.*

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In *Proceedings.Hate Speech Dataset from a White Supremacy Forum*. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings.BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Association for Computational Linguistics.

D'Sa, A. G., Illina, I., and Fohr, D. (2020). Bert and fasttext embeddings for automatic detection of toxic speech. In *Proceedings.BERT and fastText Embeddings for Automatic Detection of Toxic Speech*.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer.

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

Fortuna, P., Ferreira, J., Pires, L., Routar, G., and Nunes, S. (2018). Merging datasets for aggressive text identification. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics.

Fortuna, P., Soler-Company, J., and Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3).

Founta, A.-M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., and Leontiadis, I. (2019). A unified deep learning architecture for abuse detection. In *Proceedings.A Unified Deep Learning Architecture for Abuse Detection*, page 105–114.

Gambäck, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate-speech. In *Proceedings.Using Convolutional Neural Networks to Classify Hate-Speech*. Association for Computational Linguistics.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., and Asokan, N. (2018). All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. Association for Computing Machinery.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS One*, 14(8).

Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. In *Proceedings.Detecting Hate Speech in Social Media*.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning based text classification: A comprehensive review. *Arxiv*.

Mishra, S. and Mishra, S. (2019). 3idiots at hasoc 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages. *CEUR-WS*, 1.

Mozafari, M., Farahbakhsh, R., and Crespi, N. (2019). A bert-based transfer learning approach for hate speech detection in online social media. In *Proceedings.A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media*. Springer.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Paschalides, D., Stephanidis, D., Andreou, A., Orphanou, K., Pallis, G., Dikaiakos, M. D., and Markatos, E. (2020). Mandola: A big-data processing and visualization platform for monitoring and detecting online hate speech. *ACM Transactions on Internet Technology*, 20(2):1–21.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Lan-*

guage Resources and Evaluation, 55(2):477–523.

Ranasinghe, T., Sarkar, D., Zampieri, M., and Ororbia, A. (2021). Wlv-rit at semeval-2021 task 5: A neural transformer framework for detecting toxic spans. *arxiv*.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In *Proceedings.Deep One-Class Classification*.

Schölkopf, B. (2001). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press.

Sun, Q., Liu, H., and Harada, T. (2017). Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64:187–201.

Swamy, S. D., Jamatia, A., and Gambäck, B. (2019). Studying generalisability across abusive language detection datasets. In *Proceedings.Studying Generalisability across Abusive Language Detection Datasets*. Association for Computational Linguistics.

Tang, R., Lu, Y., Liu, L., Mou, L., Vechtomova, O., and Lin, J. (2019). Distilling task-specific knowledge from bert into simple neural networks.

Warner, W. and Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings.Detecting Hate Speech on the World Wide Web*, page 19–26. Association for Computational Linguistics.

Yin, W. and Zubiaga, A. (2021). Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*.

Zhang, Z., Robinson, D., and Tepper, J., (2018). *Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network*, chapter Chapter 48, pages 745–760. Lecture Notes in Computer Science.