Hate Speech Detection: A Deep One-Class SVM Approach,

Saugata Bose, School of Computing and Information Technology, University of Wollongong

Hate Speech Detection: not a conventional detection problem

Generally perceived as an information retrieval problem from documents (see Figure 1). In general, if a dataset contains N number of documents as D= {X1, X2, …, XN}, where Xi refers to a data point with T number of features such that Fi= {F1, F2, …, FT}. And each data point is labelled with m different classes. The classifier model is trained to identify the class of a document having class unknown.

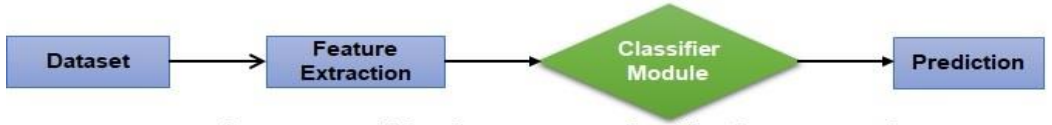
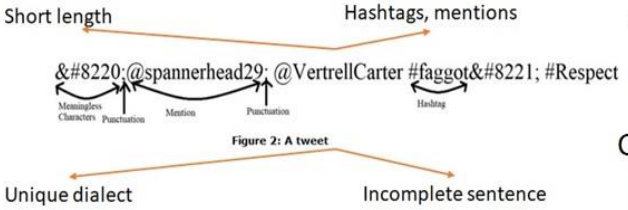


Figure 1: Traditional Document Classification Approach.

Hate speech detection from a social media post differs from traditional information retrieval system. Why?



Consider this tweet

Challenging task

- Insufficient publicly available hate speech dataset[1]
- Absence of benchmark datasets[2]
- Presence of a fuzzy boundary between hate speech and cyberbullying, abusive language, discrimination, profanity, toxicity, flaming, extremism and radicalization concepts[3].

How to Solve?

Linear [4] | Deep Learning [5] | Hybrid[6]

Is there any similarity?

Binary Classification model

- Majority of the data are non-hate in the dataset.
- 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.
- The training set does not resemble the 'true' distribution, the generalization performance is poor.

What's the problem?

What I suggest

The Classifier must know only one class and and its features: which is the hate class. The other instances will be treated as anomalies or outliers as their features deviates significantly from other observations[7].

[1] MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Go-arian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. PLoS One.
[2] Swamy, S. D., Jamatia, A., and Gamb ¨ack, B. (2019). Studying generalisability across abusive language detection datasets. In Proceedings. Studying Generalisability across Abusive Language Detection Datasets. Association for Computational Linguistics.
[3] Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys, 51(4):1–30.
[4] 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.
[5] Gamb ¨ack, 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 Computa tional Linguistics.
[6] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content
[7] Chalapathy, R., Menon, A. K., and Chawla, S. (2019). Anomaly detection using one-class neural networks. Arxiv.

Research Contributions & Proposed Model

- **One-class detection model combined with deep learning approach outperforms a binary-class detection model.**
- **Deep One-Class Hate Speech Detection Model outperforms the baseline models.**
- **Deep One-Class Hate Speech Detection Model is generalized.**

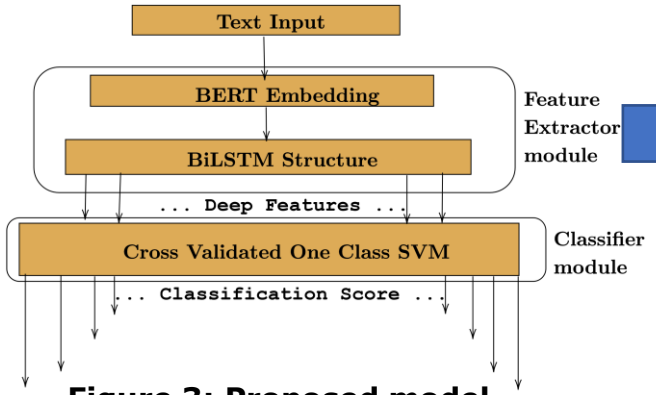
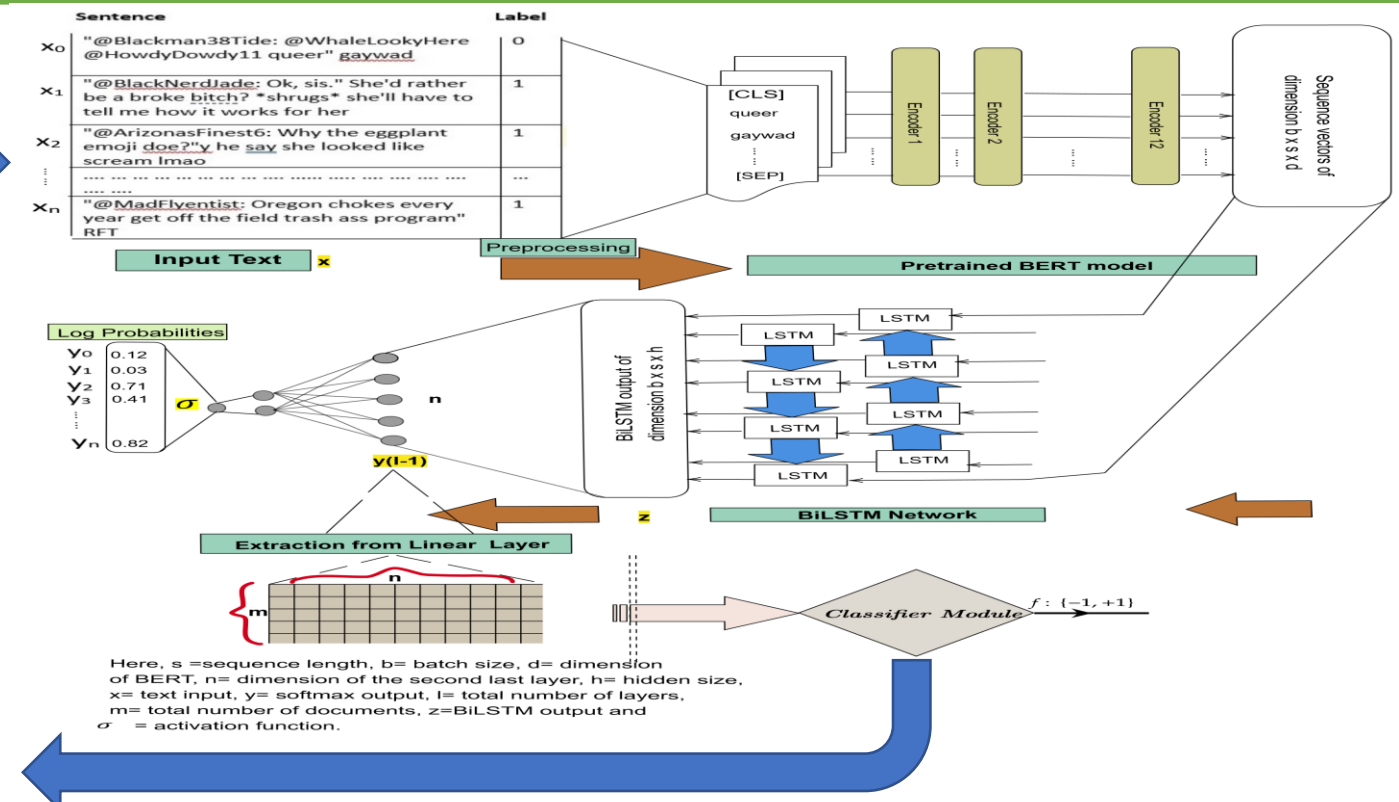
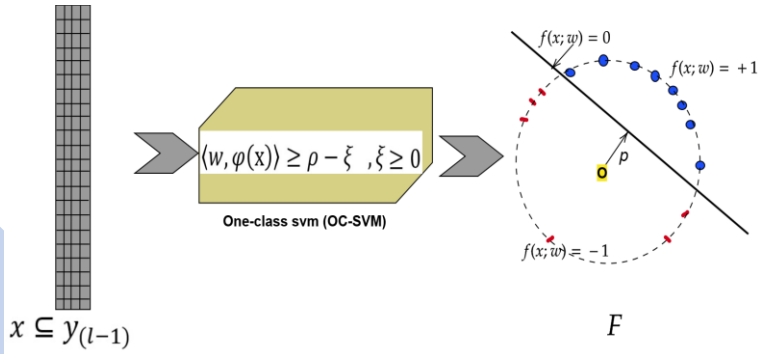


Figure 3: Proposed model





The study considers hate speech classification as a one class classification problem where the classifier will be trained only hate class features and will be able to detect anomalies during validation and test phase while the prediction differs from actual state. 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[1] in this study which is hate class.

[1] Sch¨olkopf, B. (2001). Learning with kernels : support vector machines, regularization, optimization, and beyond. MIT Press.

Experiments &Findings

Datasets: Models were evaluated with Davidson[1]*, SemEVAL-2019[2]*, Stormfront[3]*, HASOC-2019[4]* and a combination of these four datasets. English posts were considered.
Baseline methods: Proposed model was compared with CNN[5]*, LSTM, BERT-base[6]* and BiLSTM.
Performance Metric: F1 score of the hate class.

- Except CNN, each model shows improvement in one class classification approach.
- Proposed model improves performance for Stormfront, SemEval-2019 and Davidson dataset

Model / Dataset	BiLSTM			BERT-base			LSTM			CNN		
	2class	1class		2class	1class		2class	1class		2class	1class	
Stormfront	*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
Semeval	*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
Davidson	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
HASOC	*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 **Proposed Model.**

Trained with \ Tested 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.

The proposed model was used to test generalisability across the other datasets. The model improves F1 score after being trained with the mixed dataset.

*References are listed in the next slide.

Conclusion and Future plans

Through comprehensive experiments, I found that if the dataset is balanced, and if the hated class detection becomes the priority, then a one-class classifier will be a best option. I experimented with several state-of-the-art methods and with publicly available datasets. Results demonstrated that the proposed model offered the best detection and generalization results.

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

Acknowledgments

The presenter is thankful to Dr. Guoxin Su for the continuous support.

[1] 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.
[2] de Gibert, O., Perez, N., Garc´ıa-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 Linguistic.
[3] 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.
[4] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019.
[5] 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.
[6] 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 Linguistic.