Hate Speech Detection: Performance based upon a Novel Feature Detection

Saugata Bose
PhD candidate, University of Wollongong

Hate speech detection in short text on social media becomes an active research topic in recent years as it differs from traditional information retrieval for documents. I have proposed a deep learning model to improve the detection performance for deep learning using Convolutional Neural Network (CNN). Experiments have shown that it improves performance when extracted novel feature is combined with CNN and support vector machine (SVM).

Background Context

- Detecting hatred from the text is gaining attention to the researchers after few incidents happened around the globe – in Myanmar, Bangladesh and New Zealand. Facebook, Google and Twitter have been criticized for failing to deal the hated posts in their media [6].
- NLP was favorable to the researchers for classifying documents, retrieving information from documents whereas it is absolutely unsuitable for retrieving relevant features from the short texts due to the unique natures of this type of text which have mentioned previously [35].
- Non-linear classifiers are popular among researchers in solving hated text classification problem. Deep learning is gaining attention in recent years due to its automatic feature generation. And the other reason which puts deep NN ahead in this task is the volatile structure of the hated texts.

Research Contributions

- The performance of hate speech detection is highly dependent on features used to characterise the hate speech. Literature review found that several application-dependent features built by experts such as parts-of-speech etc. have been proposed. Most proposed features in the literature do not perform very well. I argue that the poor performance of those features rooted in their weak correlation with the semantics of the speech. I have proposed a feature based on hate speech lexicon. Further work will be in creating features using deep learning approach.

Methodology

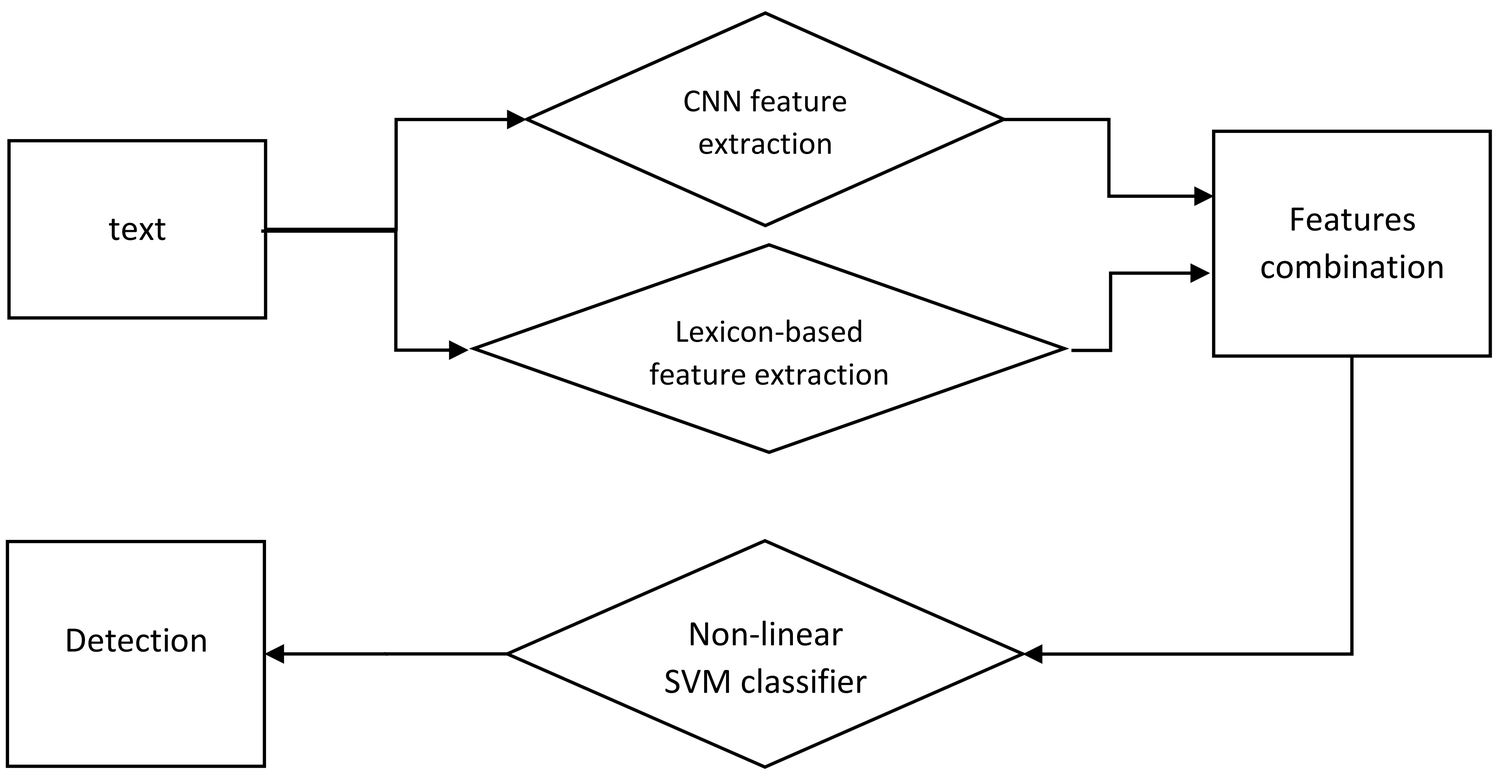

Figure 1: Proposed Framework

Experimental Setup

- I have used a publicly available corpus (24,783 number of tweets) where hate classes are 6% of the total dataset [13].
- I apply the pre trained GloVe word embedding with 100 dimensions to set the weights of our embedding layer.
- I have considered 1D convolutional layer. I assume the sequence length of the tweet to be 100 which I believe is long enough to encode tweets of any length. Then, I convert the tweets to sequences of word vectors using the word embedding, and pad or truncate the sequences to the specified sequence length.
- I use the categorical cross entropy loss function and the Adam optimiser to train the model. Here, initial learning rate of the model is 1x10-3, L2 Regularization value is 1x10-4, number of epochs are 30 and mini batch size is 150.
- I assume that if manually picked features are added to the features extracted from the trained network, it may improve the accuracy score of the classification.
- In this study, I experiment with a feature which is lexicon based. This feature will tell us how much hated or non-hate a tweet is by looking at the presence of the 'hated' and 'non hated' words. These words have strong co relation with the tweet label.
- The frequency of the hate words appearing in the hate tweets would give us a notion that how much weight a particular hated words carries in the specific tweet.
- I integrate this feature with the outputs extracted from a CNN model and fed these to an SVM classifier. Then the SVM classifier has enough features to get trained to create the margin.

Results

Hate class classification scores using CNN architecture

Dataset Type	CNN Recall Test	CNN Recall Test Hate Class	CNN Precision Test	CNN Precision Test Hate Class	CNN F1 Test	CNN F1 Test Hate Class
Balanced	0.730769	0.674825	0.692053	0.714815	0.710884	0.694245

Extracted features from CNN+ Novel feature

Dataset Type	CNN Recall Test	CNN Recall Test Hate Class	CNN Precision Test	CNN Precision Test Hate Class	CNN F1 Test	CNN F1 Test Hate Class
Balanced	0.789983	0.757986	0.727793	0.788590	0.757613	0.772985

Conclusion

- Through comprehensive experiments, I found that novel features along with extracted features from CNN outperforms CNN only performance.
- Through comprehensive experiments, I found that if the dataset is balanced, the proposed solution 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 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.

References

[1] Hern, A. (2019, April 25). MPs criticise social media firms for failure to report criminal posts. Retrieved from https://www.theguardian.com/media/2019/apr/24/mps-criticise-tech-giants-for-failure-to-report-criminal-posts-twitter-facebook-google-youtube
[2] Pitsilis G.K., Ramampiaro H. and Langseth H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence. 48(12), Pages 4730-4742. Retrieved from https://link.springer.com/article/10.1007/s10489-018-1242-y
[3] Davidson T., Warmsley D., Macy M. and Weber I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the ICWSM 2017. Retrieved from https://arxiv.org/pdf/1703.04009.pdf . Data available at https://data.world/ml-research/automated-hate-speech-detection-data