Exploring the Use of Lexicons to aid Deep Learning towards the Detection of Abusive Language

Anna Koufakou **Jason Scott**
Software Engineering Department
Florida Gulf Coast University
Fort Myers, Florida, USA
akoufakou@fgcu.edu

Abstract

Detecting abusive language is a significant research topic, which has received a lot of attention recently. Our work focused on detecting personal attacks in online conversations. As previous research on this task has largely used deep learning with word embeddings, we explored the use of sentiment lexicons as well as semantic lexicons towards improving accuracy. This is a work in progress, and our preliminary results showed promise for utilizing lexicons for this task.

1 Introduction

Detecting abusive language is a very important research topic, given the pervasiveness of social media and the surge in abusive online behavior in recent years. Online abuse has led users to quit a particular online site, move away from their home, or even commit suicide. Governments as well as social media platforms are now under pressure to detect and remove abusive users. On the other hand, online communities thrive on free speech and would be damaged by flagging and removing innocent users. At first glance, NLP models can learn linguistic patterns in conversations and detect offensive speech using features such as swear words or racial/sexist slurs. This becomes a difficult research problem as online conversational text contains casual language, abbreviations, misspellings, slang, etc. Additionally, there are gray areas which make it hard to determine if a comment is actually offensive or abusive.

Word embeddings have been used successfully in many NLP tasks such as sentiment analysis or classification. The current state-of-the-art in the field of abuse detection in online conversations is based on deep learning with word embeddings; for example, see (Gamback and Sikdar, 2017; Gunasekara and Nejadgholi, 2018; Zhang et al., 2018) among others. In this study, we explored the application of lexicons with word embeddings towards the task of detecting and classifying abusive language. Specifically, we applied convolutional neural networks to automatically identify comments which contain abusive language. Our research explored the use of sentiment lexicons, a form of sentiment dictionary associating words with sentiments. We also explored enhancing the word embeddings with semantic lexicons, which contain semantic relationships between words (for example, synonyms).

2 Experiments

2.1 Data Description

We first acquired and pre-processed text comments from the Wikimedia Toxicity Project, a project which has released datasets of English text comments from the editor's forum of the platform (Wulczyn et al., 2017). The resulting dataset contains 115,841 text comments, each with annotations by 10 human workers which indicate whether or not each worker believed the comment contained a personal attack. The data also contains additional fields such as the type of attack; we used only the comment text and if it contained an attack or not. The dataset was divided into comments containing an attack and comments

not containing an attack using the human annotations. A comment was labeled as containing an attack if a majority of human workers annotated it as an attack. We focused on the entire data: 87.7% of the comments in the data do not contain a personal attack, while 12.3% of the comments do.

2.2 Experimental Setup and Preliminary Results

We developed baseline experiments and compared several variations of combining lexicons with word embeddings for detecting attacks in the data. We first pre-processed the data and generated Word2vec embeddings from the text (dimension of 200). As our baseline, we employed a three-layer convolutional neural network (CNN) (Kim, 2014). We also experimented with recurrent neural networks (LSTMs, GRUs), and fastText (Bojanowski et al., 2016); we presented the top results for comparison purposes.

We then experimented with Lexicon Integrated CNN models (Shin et al., 2017). These ideas involved creating sentiment embeddings from sentiment lexicons and then integrating the sentiment embeddings to the baseline CNN. We only employed naïve concatenation and parallel convolution due to time constraints. We used the following sentiment lexicons: AFINN-96, MSOL-June15-09, Bing-Liu Opinion, and NRC EmoLex. Secondly, we explored enhancing the word embeddings by "retrofitting" them to semantic lexicons as proposed by Faruqui et al. (2015). We used the PPDB-XL semantic lexicon as it was shown to have the best performance in previous research.

For our implementation, we used Python, scikit-learn, and TensorFlow executed on a Google Cloud TPU on the TensorFlow Research Cloud, using a free trial on the TPU[1]. We evaluated the network after 10,000 TPU steps of training with a randomly shuffled and batched training dataset, a learning rate of 0.001, dropout of 0.5, Adam optimizer, 10-fold cross validation, and 80-20 training-test split. The LSTM/GRU/fastText experiments were run on a single-GPU machine using Keras[2] with 10 epochs, 128 batch size and various learning rates (we reported the best results we observed).

Table 1 shows our preliminary results. We reported accuracy, F1-score, recall and precision. Besides the CNN experiments, we included representative experiments for a bidirectional GRU, fastText, and an LSTM with fastText embeddings for comparison. "Retrofitting" word embeddings to semantic lexicons (Faruqui et al., 2015) had the best performance overall. The two techniques from (Shin et al., 2017) did not have a difference from the baseline. More research is needed for these ideas: for example, it might be helpful to use many different lexicons, as well as attention mechanisms.

3 Conclusions

We explored the use of lexicons, semantic or sentiment, towards the detection of personal attacks in online conversations. Due to time constraints and lack of appropriate infrastructure, we showed results for a few techniques and lexicons. Our experiments so far provide evidence that enhancing word embeddings using semantic lexicons (Faruqui et al., 2015) shows more promise. As plans are underway for our department to acquire appropriate infrastructure, we aim to experiment more with integrating sentiment lexicons into the CNN, as well as with more recent works by Mrkšić et al. (2017) or by Glavaš and Vulić (2018). We also plan to explore hate speech lexicons such as in (Gitari et al., 2015).

Embeddings	*Model*	*Accuracy*	*F-1 Score*	*Precision*	*Recall*
Word Embeddings	Bidirectional GRU	0.94	0.74	0.83	0.61
N/A	fastText	0.93	0.78	0.85	0.47
fastText Embeddings	LSTM	0.94	0.75	0.84	0.69
Word Embeddings	CNN (Baseline)	0.96	0.82	0.84	0.79
Sentiment + Word Embeddings (Shin, 2017)	Naïve Concatenation CNN	0.96	0.82	0.83	0.81
	Parallel CNN	0.96	0.81	0.86	0.77
"Retrofitted" Word Embeddings (Faruqui, 2015)	CNN	**0.98**	**0.90**	**0.95**	**0.84**

Table 1: Preliminary classification results comparing word embeddings versus integrating lexicons

[1] TensorFlow Research Cloud https://www.tensorflow.org/tfrc. We received a free trial which was used for all experiments.
[2] https://keras.io

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(135–146).

Manaal Faruqui, Jesse Dodge, Sujay Jauhar, Chris Dyer, Eduard Hovy, and Noah Smith. 2015. Retrofitting word vectors to semantic lexicons. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics NAACL*.

Bjorn Gamback and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. *Proceedings of the First Workshop on Abusive Language Online, ALW1*.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Goran Glavaš, and Ivan Vulić. 2018. Explicit retrofitting of distributional word vectors. *Proceedings of the Annual Meeting of the Association for Computational Linguistics ACL*.

Isuru Gunasekara, and Isar Nejadgholi. 2018. A Review of Standard Text Classification Practices for Multi-label Toxicity Identification of Online Content. *Proceedings of the 2nd Workshop on Abusive Language Online ALW2.*

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP.*

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó. Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Proceedings of the Annual Meeting of the Association for Computational Linguistics ACL.*

Bonggun Shin, Timothy Lee, and Jinho Choi. 2017. Lexicon Integrated CNN Models with Attention for Sentiment Analysis. *Proceedings of the EMNLP Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. *Proceedings of the 26th International Conference on World Wide Web WWW*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. *European Semantic Web Conference*.