Offensive Speech Detection on Twitter

Taurean Parker, Caroline Roper

New York University

Objectives

1. To label tweets as containing hate speech, obscenity, threats, prejudice, or other offensive rhetoric.
2. To examine the effectiveness of transfer learning methods for classification problems with limited labeled data.
3. To support NYU's Social Justice Lab's research on the Ideological-Conflict hypothesis.

Introduction

This research supports NYU's Social Justice Lab within the Department of Psychology. Ultimately, the model will produce labels that will help researchers test the "Ideological-Conflict Hypothesis" which suggests that liberals and conservatives express similar levels of intolerance toward 'ideologically dissimilar and threatening groups'[1].
Our phase of the research seeks to identify an optimal method for detecting hate speech, prejudice, and offensive language on Twitter using a dataset of 7k labeled tweets. The best model will be applied to a dataset of 734k tweets and cross-referenced with the political ideology of the Twitter user to evaluate the Ideological-Conflict Hypothesis.

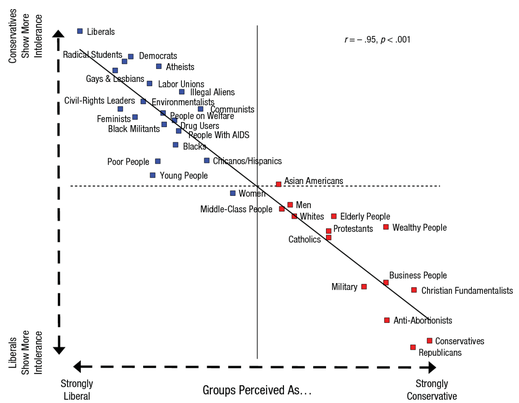

Figure 1: Representation of Ideological-Conflict Hypothesis[1]

Label Definitions

Label	Definition
Stereotypes	False or misleading generalizations about groups held in a manner that renders them largely immune to counterevidence
Pornography	Explicitly sexual material intended for sexual stimulation
Non English	Tweets written in non-English
Negative Prejudice	An aversive or hostile attitude toward a person who belongs to a group, simply because he belongs to that group
Name Calling	Abusive or insulting language referring to a person or group
Hate Speech	Communication that carries no meaning other than the expression of hatred for some group
Threat	The intentional use of physical force or power, threatened or actual
Obscenity	Offensive word or phrase considered inappropriate to use in professional settings

Table 1: Label definitions, abbreviated from annotator coding manual

Related Work

SVM - Has been successfully implemented to detect hate speech in prior work[2].
LSTM - Our model was inspired by the work of Badjatiya et al. which also used an LSTM with GloVe embeddings to detect hate speech on Twitter[3].
ULMFiT (Universal Language Model Fine-Tuning)[4] - Consists of three stages:

1. General-domain language model pre-training
2. Domain-specific language model fine-tuning
3. Target task classifier training

BERT (Bidirectional Transformers for Language Understanding) - Developed by researchers at Google, BERT is a language understanding system novel for its bi-directional architecture and unique pre-training tasks[5].

Methods

The table below summarizes which aspects of each model we performed experiments on.

	SVM	LSTM	ULMFiT	BERT
Text Pre-Processing	✓	✓		✓
Regularization	✓		✓	
Vocabulary Size	✓	✓	✓	
Batch Size		✓	✓	✓
Seq Len		✓	✓	
Early Stopping		✓		
Learning Rate			✓	✓
Number of Epochs			✓	✓

The SVM uses the tf-idf weights of n-grams from unigrams to 5-grams.
The LSTM has one recurrent layer, uses dropout for regularization, and is initialized from GloVe vectors:

1. Vectors pre-trained on 2B tweets
2. Our own vectors on all 734k tweets in our dataset

Below are select results of LSTM parameter tuning, predicting hate speech:

batch size	hidden	seq leng	vocab	f-score
32	400	40	12k	0.750
16	400	40	12k	0.744
32	200	40	12k	0.696
32	200	40	10k	0.682
32	200	30	10k	0.667
32	200	25	10k	0.640

Results

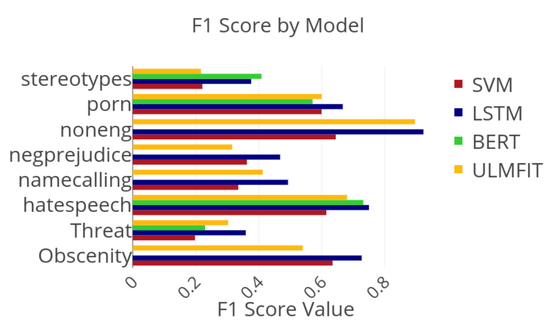

Figure 2: SVM, LSTM, BERT, and ULMFiT validation f-scores

Conclusions

1. The LSTM model performed the best.
2. Some labels such as hate speech, obscenity, and pornography, could be predicted very reliably.
3. Other labels such as negative prejudice, stereotypes, name-calling, and threat were considerably harder to predict. The original data annotators exhibited higher disagreement on these labels.
4. Transfer learning methods ULMFiT and BERT didn't outperform the LSTM. Twitter data differs greatly from the general domain datasets that the models are trained on.

References

[1] Mark J. Brandt, Christine Reyna, John R. Chambers, Jarret T. Crawford, and Geoffrey Wetherell. The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science*, 23(1):27–34, 2014.

[2] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[3] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[4] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *ACL*. Association for Computational Linguistics, 2018.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

Acknowledgements

We would like to thank Vivienne Badaan and Mark Hoffarth from the social psychology department for their expertise in providing the initial 7K tweets

Web

- https://github.com/NYU-CDS-Capstone-Project/TwitterHateSpeechDetection