Interpretable Multi-Modal Hate Speech Detection

Prashanth Vijayaraghavan[1]; Hugo Larochelle[2]; Deb Roy[1]
[1]MIT Media Lab, [2]Google Brain

Abstract

With growing role of social media in shaping public opinions and beliefs across the world, there has been an increased attention to identify and counter the problem of hate speech on social media. Hate speech on online spaces has serious manifestations, including social polarization and hate crimes. While prior works have proposed automated techniques to detect hate speech online, these techniques primarily fail to look beyond the textual content. Moreover, few attempts have been made to focus on the aspects of interpretability of such models given the social and legal implications of incorrect predictions. In this work, we propose a deep neural multi-modal model that can: (a) detect hate speech by effectively capturing the semantics of the text along with socio-cultural context in which a particular hate expression is made, and (b) provide interpretable insights into decisions of our model. By performing a thorough evaluation of different modeling techniques, we demonstrate that our model is able to outperform the existing state-of-the-art hate speech classification approaches. Finally, we show the importance of social and cultural context features towards unearthing clusters associated with different categories of hate.

Dataset

Datasets	Details
[3]	None: 53.8%; Hate: 4.96%; Abusive: 27.15%; Spam: 14%; Tweets: ~100k
[4]	None: 16.8%; Hate: 5.8%; Offensive: 77.4%; Tweets: ~25k
[5]	None: 68%; Sexism: 20%; Racism: 11%; Tweets: ~18k
[6]	None: 74%; Harassment: 26%; Tweets: ~21k
Our Dataset	None: 58.1%; Hate: 16.6%; Abusive: 25.3% Tweets: ~258k

Tweets

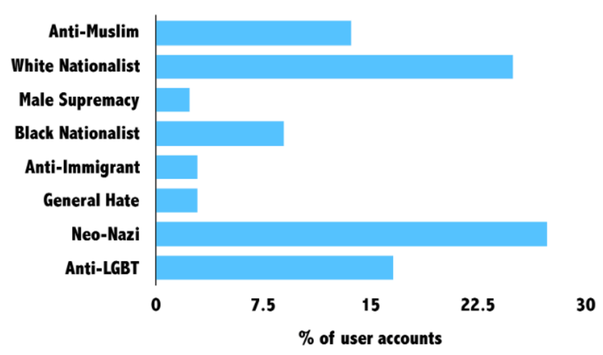
In order to expand our dataset, we perform an exploratory search on Twitter that consists of phrases containing:

(a) swear words combined with positive adjectives (eg. “f**king awesome”, “damn good”, “bloody wonderful”)

(b) swear words combined with races, religions, sexual orientations or derogatory references to them. (e.g. “f**king ragheads”, “sh**ty muslims”)

Hate groups

We gather the extremist groups data from Southern Poverty Law Center (SPLC) and map these groups on Twitter. We got ~3k user accounts containing such information and filtering for inactive accounts. We construct a directed graph *G* where each vertex is a user and edges represent their relationships (friends & followers). We compute the page rank of this graph *G* and obtain the top ~10k accounts including the ~3k seed user accounts.



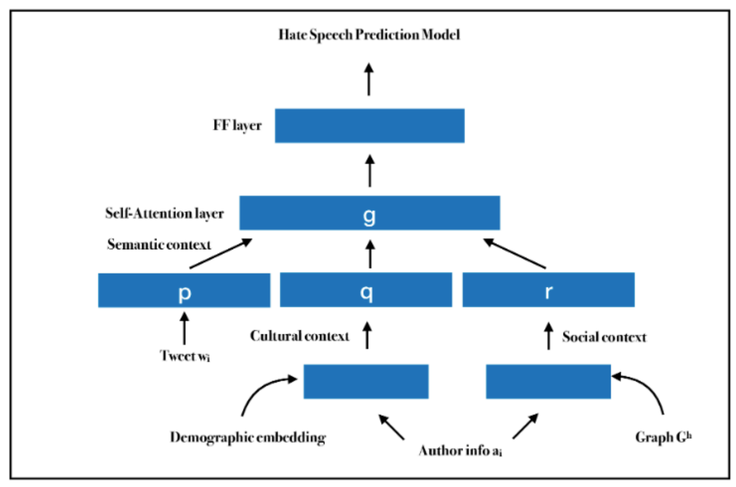
Hate Category	Sample user accounts
Anti Muslim	@clarionproject, @PoliticalIslam
White nationalist	@redicetv, @IdentityEvro_IN
General Hate	@JewishTaskForce, @hatreon
Male Supremacy	@ReturnOfKings, @Bashflak
Black Nationalist	@ArmyOfIsrael, @DrAvaMuhammad
Anti-Immigrant	@FAIRImmigration, @CIS_org
Neo-Nazi	@GoldenDawnHOGD, @GDWatch_EN
Anti-LGBT	@savecalifornia, @AmericanFamAssc

Model

The data includes tweets, author attributes and social structure represented by author's followers and friends. Let us denote our hate dataset as $D^{(H)} = \{(w_1, a_1),(w_2, a_2), ..., (w_n, a_n,)\}$, where each tuple in this set consists of the tweet text w_i, author information a_i used to derive social and cultural context associated with the tweet. Defining the input as $x_i=(w_i, a_i)$, we denote our model as:

$$f_\vartheta(x_i) \approx g_\vartheta(P(w_i), Q(a_i), R(a_i))$$

where *P, Q, R* are neural architectures extracting semantic features from tweet text (*P*) [1] and socio-cultural features(*R, Q*) [2] from author information; *g* is the function that determines the fusion strategy.



Evaluation

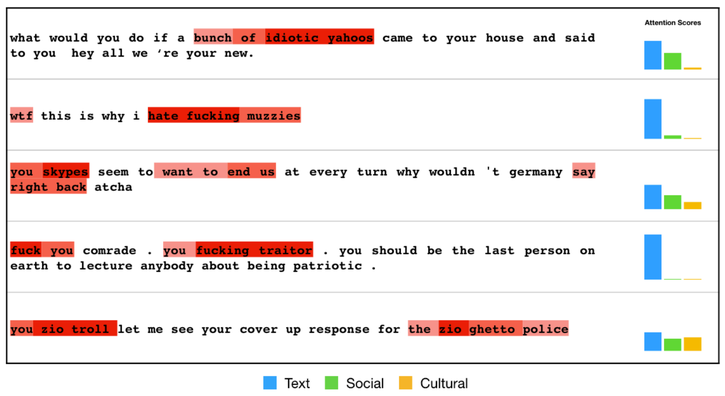
- Tables show that social and cultural context improves the performance of our model significantly compared to purely state-of-the-art text-based models.
- The results indicate that the model that uses social and cultural context is able to produce better clusters having more overlap with good hate categories compared to the text only model.

Model	F1 (Hate)	F1 (Overall)
Traditional Models: Text+Social+Demographic		
LR	0.53	0.72
SVM	0.563	0.729
DL Models: Text Only		
CNN-Char	0.735	0.866
BiGRU+Char+Attn	**0.744**	**0.864**
CNN-Word	0.658	0.788
BiGRU+Attn	0.683	0.801
BiLSTM-2DCNN	0.661	0.795
DL Models: Text+SC		
CNN-Char+FF	0.760	0.879
BiGRU+Char+Attn+FF	**0.784**	**0.90**

Top 5 words	Hate Category
jihadi, muzzie, terrorist, #stopislam, #banmuslim	Anti-Islam
n**ga, n**ger, #whitepower, ghetto #14words	Anti-Black
#buildthewall, illegals, #noamnesty, #illegalaliens, #anchorbabies	Anti-Immigrant
f**k, c**t, hate, b**ch, a****le	General Hate
#antisemitism, #antisemites, nazi, satan, neonazi	Anti-semitic

Interpretability

We interpret the results by highlighting words in the text and constructing bar graphs that indicate the relevance of each of the features: textual, social, cultural.



Tweets containing code words like “skypes”, “yahoos”, “zio”, “zog”, etc. attacking particular group of people have higher attention scores for social and cultural context vectors. The model is able to understand these code words and tag them as hateful content.

Conclusions

- We developed a comprehensive model to automatically detect hateful content.
- We adopt different feature extraction strategies for different modalities of data: text, demographic information and social graph.
- We derive important insights about our model and its ability to understand hate speech code words and cluster into different categories of hate speech.

Contact

Prashanth Vijayaraghavan
MIT Medial Lab
Email: pralav@mit.edu
Website: http://www.mit.edu/~pralav/index.html

References

1. Chen, Huadong, et al. "Combining Character and Word Information in Neural Machine Translation Using a Multi-Level Attention." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018.
2. Vijayaraghavan, Prashanth*, Soroush Vosoughi*, and Deb Roy. "Twitter demographic classification using deep multi-modal multi-task learning." Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vol. 2. 2017.
3. Founta, Antigoni Maria, et al. "Large scale crowdsourcing and characterization of twitter abusive behavior." *Twelfth International AAAI Conference on Web and Social Media*. 2018.
4. Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." *Eleventh International AAAI Conference on Web and Social Media*. 2017.
5. Golbeck, Jennifer, et al. "A large labeled corpus for online harassment research." *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 2017.
6. Park, Ji Ho, and Pascale Fung. "One-step and two-step classification for abusive language detection on twitter." *arXiv preprint arXiv: 1706.01206* (2017).