Neural Content Moderation

By Isabella Garcia-Camargo, Martin Amethier, Guy Wuollet

Motivation

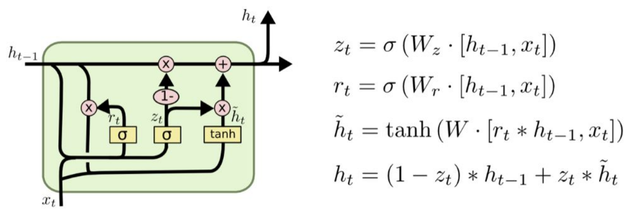
- Content Moderation is crucial for healthy online social spaces
  - Different communities have different standards
  - Content volume and moderator mental health requires automated moderation

Data

We have two datasets. The first is a Kaggle dataset from the *Toxic Comment Classification Challenge* and contains Wikipedia comments. The second is Reddit data and comes from Pushshift and from *Hybrid Approaches to Detect Comments Violating Macro Norms on Reddit.* The Reddit combines 2 million moderated comments and 2 million unmoderated comments from top 100 subreddits by popularity during 2016-2017.

	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	Thank you for understanding. I think very high...	0	0	0	0	0	0
1	:Dear god this site is horrible.	0	0	0	0	0	0
2	"::: Somebody will invariably try to add Relig...	0	0	0	0	0	0
3	" \n\n It says it right there that it IS a typ...	0	0	0	0	0	0
4	" \n\n == Before adding a new product to the l...	0	0	0	0	0	0
...	...	...	...	...	...	...	...
145	Simple: You are stupid!	1	0	1	0	1	0
146	The only group the US is unambiguously support...	0	0	0	0	0	0
147	" \n\n \n :Oh, that's because they don't. Or ...	0	0	0	0	0	0
148	" \n\n ::: You have my trust. But trust me on ...	0	0	0	0	0	0
149	" \n\n ==BLP== \n I've removed some contentiou...	0	0	0	0	0	0

Image 3: LSTM Cell Architecture



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Image 1: GRU Model Architecture

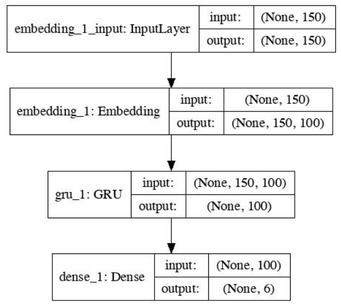
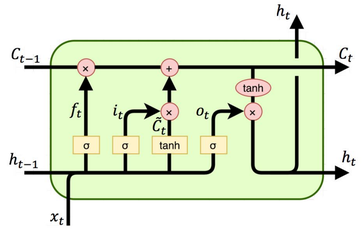


Image 2: GRU Cell Architecture



Looking Forward

- Data Augmentation is a great next step. We built transformation functions to swap adjectives and replace nouns, verbs, and adjectives with synonyms. Unfortunately, Snorkel's SSL certificate failed and we were unable to troubleshoot.
- As different communities have different standards, it would be useful to train classifiers on specific subreddit data. This could be done using transfer learning from a more general model, or as a framework for creating domain specific models.

Models

- We ran a total of seven models: LSTM, GRU, Double-LSTM, Double-GRU, Triple-LSTM, Bidirectional-LSTM, and a TF-IDF NN.
- We found that the GRU outperformed the LSTM models. This is likely due to the shortness of our sequences, which were on average 300 tokens long with a median of 150 tokens.
- Bidirectionality also improved our results, signaling the importance of backwards relationships between tokens.

Table 1: Wikipedia Comment Results

	Train Accuracy	Test Accuracy	Δ in Accuracy
MODELS			
LSTM	0.9854	0.9810	.0044
Two Layer LSTM	0.9855	0.9807	.0048
Three Layer LSTM	0.9853	0.9795	.0058
GRU	0.9883	0.9814	.0069
Two Layer GRU	0.9869	**0.9816**	.0053
TF-IDF NN	**0.9994**	0.9743	**.0251**
Bidirectional LSTM	0.9864	0.9810	.0054
BASELINES			
TF-IDF Logistic	0.9577	0.9543	.0034
Naive Bayes SVM	0.8899	0.8745	.0154

Table 2: Reddit Comment Results

	Train Accuracy	Test Accuracy	Δ in Accuracy
TF-IDF NN	0.7631	0.6758	0.873

References

[1]https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona
[2] https://www.snorkel.org/
[3] https://www.kaggle.com/jhoward/nb-svm-strong-linear-baseline
[4] Chandrasekharan, Eshwar, Samory, Mattia, & Gilbert, Eric. (2019). Hybrid Approaches to Detect Comments Violating Macro Norms on Reddit (Version 2.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3338698
[5] Hamilton, William L., et al. (2016) Inducing domain-specific sentiment lexicons from unlabeled corpora. {\it Proceedings of the Conference on Empirical Methods in Natural Language Processing.} Vol. 2016. NIH Public Access.