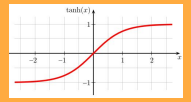
Offensive tweet detection is inherently biased.

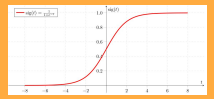
SemEval ‘19- OffenseEval Competition, Sub-task A: Detecting Offensive Tweets

By Emmi Bevensee

Methods:

- BiDirectional GRU (128 Nodes) Layer
- Global Max Pooling
- Unique Words Vocab
- Binary Cross-Entropy Loss
- Tanh Activation on hidden layer
- Sigmoid output





Results:

Accuracy: 0.796512
F1-micro: 0.796512
F1-macro: 0.729102
F1-weighted: 0.788812

State of the art models using BERT, additional datasets, and teams of NLP professionals still only yield around an 80% F1-Macro score. My model is competitive with other RNNs such as the BiLSTM baseline at 75%.

Loss and Accuracy on Dev Set:

Stopped at 2 epochs to prevent overfitting.

Hypothesis:

Even competitive Recurrent Neural Networks trained on hand-annotated data will have a high-degree of errors in detecting offensive content as a result of biased datasets.

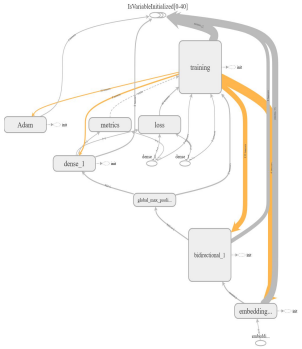
Details of Task:

- 13,240 “gold-labeled” training Tweets
- Labeled as Offensive or Not Offensive
- 860 Test examples collected differently than training data to symbolize transfer learning.

Conclusions

Neural networks just learn whatever we teach them. Although it is necessary to continue testing technical solutions to the problem of harmful content detection, there is an inherent bias to the task itself. Because “offensiveness” is subjective, we must rely on greater transparency and expert political guidance in navigating complex and controversial topics surrounding harmful content moderation such as offensiveness detection. Platforms are always making choices about what content they are willing to host, therefore it is important to push them in ethical directions rather than solely pursuing a utopian technical solution to a problem that is largely experiential.

Computational graph of my model



Error Analysis

My model says it’s offensive but the labels say it’s not:
“he probably gets paid to say that...with $$ and assurances that he won't be called an islamophobe---pfffft it's a fake word--justin is an IDIOT and he is destroying canada one refugee at a time”

- This does seem like it would be labeled offensive. Shows coding variability.

My model says it’s not offensive but the labels think it is:
“thank you sm and it was great meeting you too!! seeing you and my other favs reminded me that its so worth pushing through the pain to be able to do things like that!!”

- This is clearly mislabeled as offensive.

Tweets that only a political analysis would reveal as designed to cause offense but are labeled as not offensive:
“I like my soda like I like my boarders [sic] with a lot of ICE.”

- This is an intentional verbal attack against undocumented persons.

Alternative Strategies

Aside from just using a different form of embedding (ie. GloVe, Word2Vec) or fine-tuning entirely on BERT, the results of my model, the state of the art, and the kinds of errors resulting from biased labels show the need for political awareness and quality control in the creation of gold-standard labels in hand-annotated datasets.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). *SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)*. Retrieved from https://arxiv.org/abs/1903.08983v2