Toxic Comment Classification and Unintended Bias

Chady Ben Hamida, Victoria Ge, Nolan Miranda

{chadybh, vge, mirandan}@stanford.edu

Motivation

- Internet users find it much easier to propagate harmful stereotypes and toxic commentary in comment sections
- Unfortunately, this promotes an unhealthy environment online, and toxic commentary often ropes in language regarding minorities to construct insults

Problem Statement

- Construct a model that can accurately classify toxicity of unseen comments
- Find the bias with respect to the mention of certain identities

Data

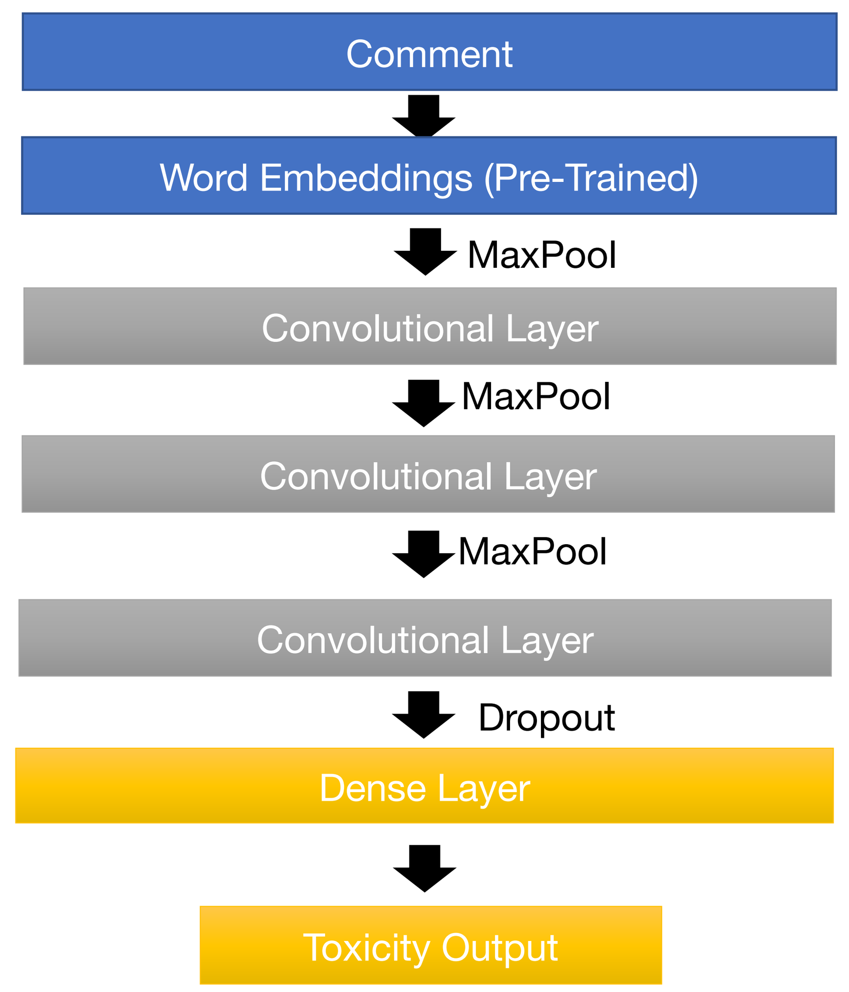
- ~1.8 millon data points taken from Civil Comments platform in 2017 hosted by Kaggle
- Each data point contains comment text and a classification label (from "very toxic" to "not toxic") across multiple graders aggregated into a score of 0.0 to 1.0
- Data points also binarize references to mentions of identity (e.g. "homosexual_gay_or_lesbian", 'christian', black')

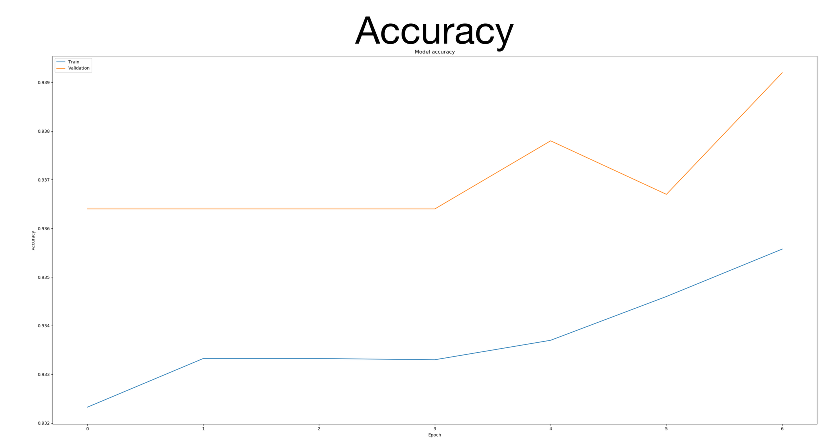
Models

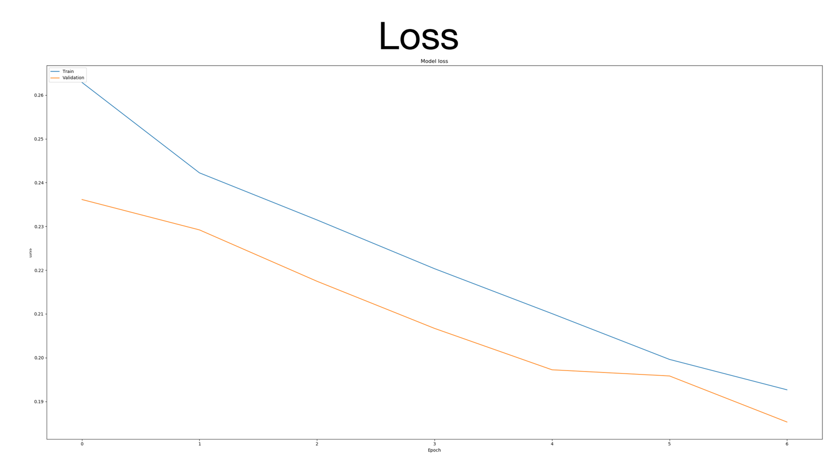
- **Naive Bayes** with Laplace smoothing: multivariate bernoulli model for bag-of-word count
- **Logistic regression**: Minimize cross-entropy loss with regularization and normal cost function on tf-idf
- **CNN**: used keras to create CNN of three convolutional layers. Interlaced with three MaxPooling layers to reduce dimensionality, speed up runtime, and to mitigate overfitting
- Employed dropout as a method of regularization and measured loss using categorical cross-entropy (softmax + cross-entropy)
- Ran each CNN for 7 epochs with batch size 128

Results

Train/Validate split: 80/20 (1443899 train, 360975 validate comments)







identity (id)	bnsp_auc	bpsn_auc	id_auc	id_size
black	0.961	0.747	0.801	3025
white	0.961	0.757	0.807	5047
homosexual_gay...	0.957	0.771	0.807	2210
muslim	0.954	0.802	0.830	4204
female	0.936	0.868	0.865	10804
...mental_illness	0.957	0.835	0.869	954
jewish	0.941	0.854	0.871	1570

Model	Embeddings	Accuracy
Naïve Bayes	(Bag-of-words)	0.9202
Logistic Reg	(TF-IDF)	0.9474
CNN	GloVe 50d	0.9390
CNN	GloVe 100d	0.9450
CNN	fastText 300d	0.9484

Features

- Bag of words: vectorizes comments in context of the total vocabulary to model term frequency
- TF-IDF: vectorizes comments in context of the total vocabulary to model relative term importance
- GloVe and fastText: downloaded pre-trained embeddings for more compact vectors gathered from complex NNs
- Identities: binarized mentions of various identity markers

Discussion

- Pre-trained word embeddings added comparably high improvements in accuracy, but higher dimensions increased runtime for little gain
- Our best CNN performed with 94.84% accuracy with word embeddings of 300 dimensions (fastText)
- Logistic regression performed very well, with a higher accuracy than a few of the CNN models
- While complex models can learn the problem decently, the simple model (logistic regression) ran about 10 times faster with similar results
- bnsp_auc (Background Negative, Subgroup Positive) and bpsn-auc (Background Positive, Subgroup Negative) are both metrics for unintended bias

Future

- We will consider improving the accuracy of our algorithm by using recurrent neural networks, namely, BLSTM which would allow us to integrate both past and future context in our model
- We can consider translating our comments to other languages then back to English as a way to increase the dataset and make it more generalized

References

[1] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and Mitigating Unintended Bias in Text Classificatio." Feb. 2018.
[2] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation." Oct. 2014.
[3] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, et al., "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech." Dec. 2015.