[Jigsaw Unintented Bias in Toxicity Classification]

[1.Ayush Pandey 2.Ahmad Alshsaref 3.Nidhi Chaudhary 4.Tanya Sharma]
[Dr. Suneet Gupta]

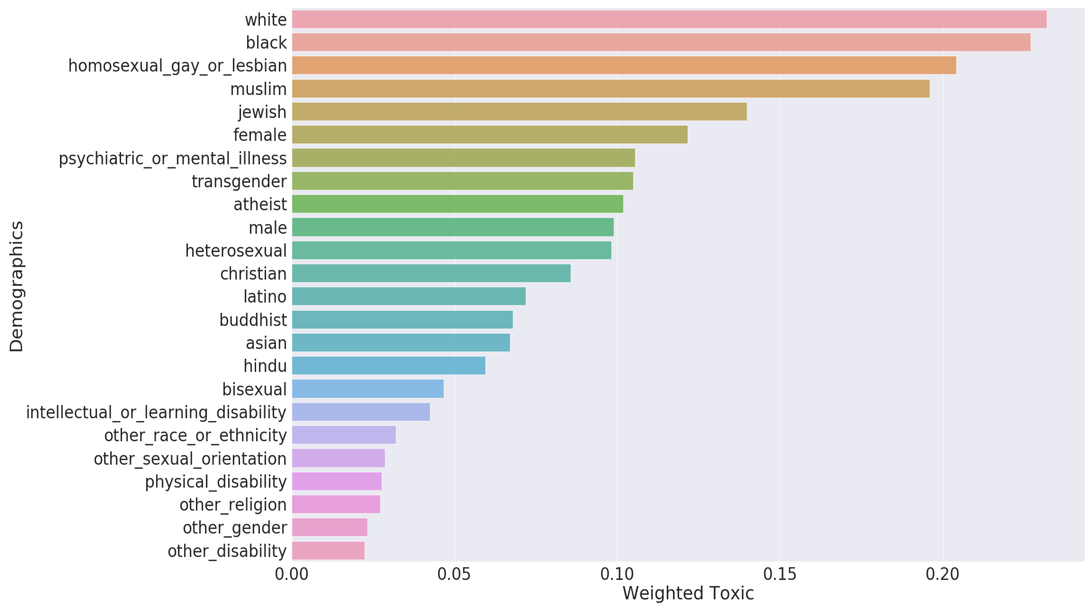
Abstract

Toxic comment classification has become an active research field with many recently proposed approaches. However, while these approaches address some of the task's challenges others still remain unsolved and directions forfurther research are needed. To this end, we compare different deep learning and shallow approaches on a new, large comment dataset and propose an ensemble that outperforms all individual models. Further, we validate our findings on a second dataset.

Introduction

In this competetion our challenge is to build a model that recognizes toxicity and minimizes this type of unintended bias with respect to mentions of identities. And here the main area of focus is machine learning models that can identify toxicity in online conversations, where toxicity is define as anything rude,disrespectful or otherwise likely to make someone leave a discussion.

Here we will be using a dataset labeled for identify and mentions and optimizing a metric designed to measure unintended bias.
And here we develop strategies to reduce a uintended bias in machine learning models.



Percent of toxic comments related to different identities, using target and popolation amount of each identity as weights

Proposed Method

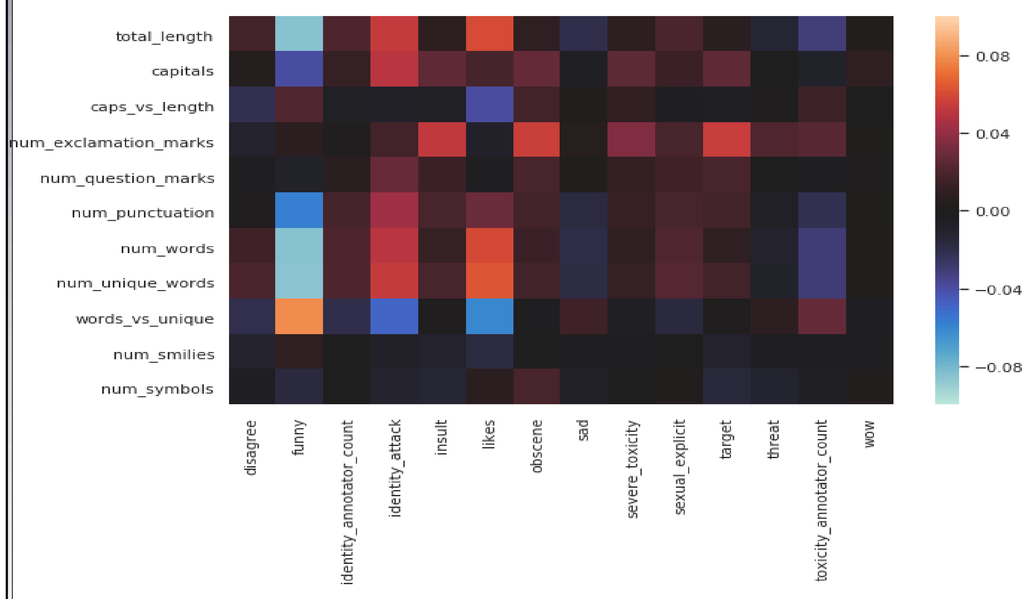
Here we used the combined model i.e., **BERT+LSTM**, which give the accuracy of **0.9417**.

LSTM(Long Short Term Memory)-: It is a special kind of RNN,capable of learning long-term dependencies. LSTM are explicitly designed to avoid the long-term dependency problem.Remembering information for long periods of time is practically their default behavior, not something they struggle to learn.

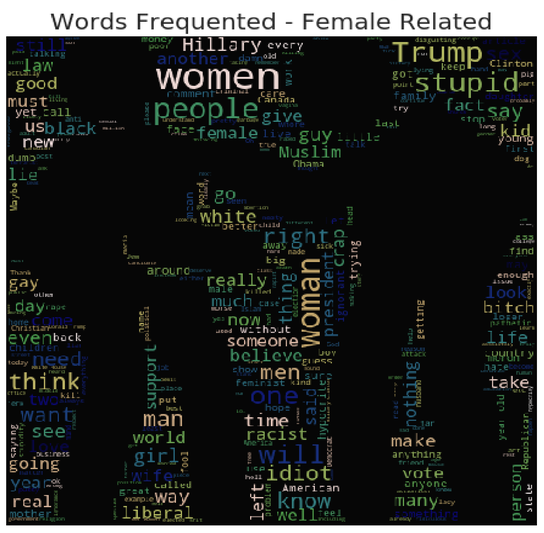
BERT(Bidirectional Encoder Representations from Transformers)-: It is a key technical innovation is applying bidirectional training of Transformer,a popular attention model to a language modelling.it has caused a stir in the machine learning community by presenting state-of-the-art results in a wide variety of NLP tasks.

Experimental Results and Discussion

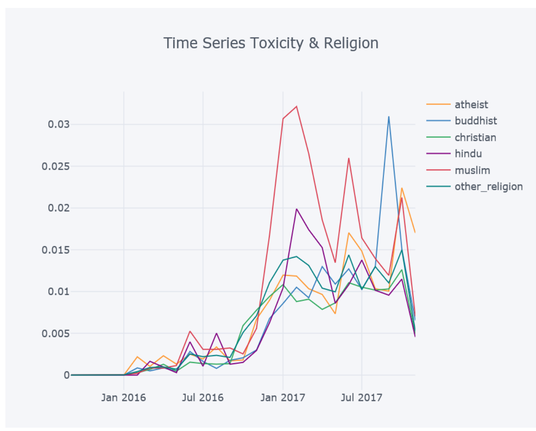
❑ Correlations between new features and targets in heatmap:

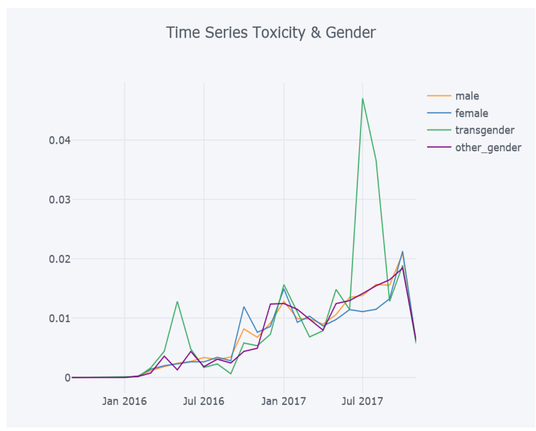


❑ Moreover, we can do something fun, digging into the text with WordCloud. Let's check the Words frequented Toxic Comments:

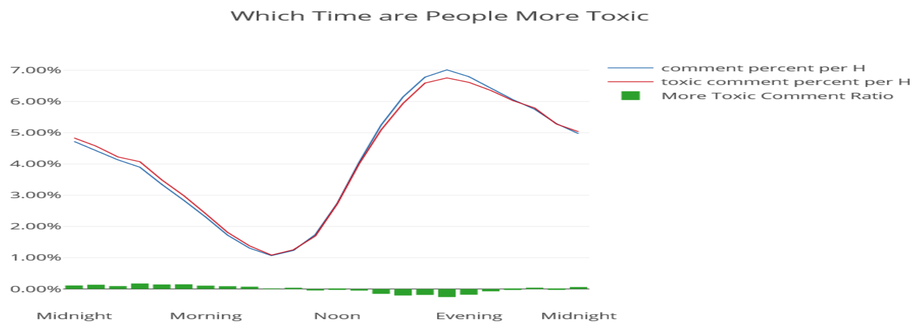


❑ Time Series Toxicity with Race, Religion, Sexual orientation, Gender and Disability





❑ Which Time are People More Toxic?



Conclusions

In this work we presented multiple approaches for toxic comment classification. We showed that the approaches make different errors and can be combined into an ensemble with improved F1-measure.

References

1. https://www.kaggle.com/thousandvoices/simple-lstm
2. https://github.com/google-research/bert

Team No. - vii