Institute for Artificial Intelligence
Franklin College of Arts and Sciences
UNIVERSITY OF GEORGIA

# SENTIMENT ANALYSIS BASED LANGUAGE MODEL EVALUATION

Aashish Yadavally, Anirudh K. M. Kakarlapudi, Sumer Singh

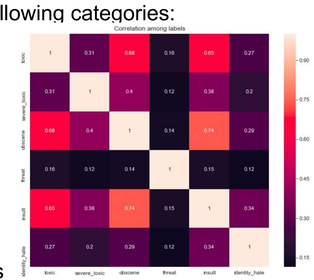
## INTRODUCTION

Language models assign probabilities to word sequences, and help guide and constrain the search among alternative word hypotheses. Thus, language modeling is an important sub-task of a majority of natural language processing (NLP) problems. With the advent of the internet and high performance computer hardware, statistical NLP techniques have become highly effective. In this work, we performed an extrinsic evaluation of different language models ranging from the basic n-grams to the recently proposed language representation model BERT [1]. We embedded these models to the downstream task of sentiment analysis on the Toxic Comment Dataset [2].

## DATASET

The "Jigsaw Toxic Comment Classification Challenge" dataset consists of Wikipedia comments that are labelled by human raters for toxic behavior belonging to the following categories:

(a) toxic,
(b) severe toxic,
(c) obscene,
(d) threat,
(e) insult,
(f) identity hate.




**Objective:**
Build a language model that predicts the toxicity of each comment.

**Challenges:**

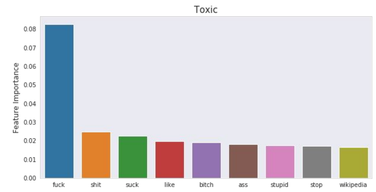
- Each comment can belong to more than one of the mentioned categories
- Exploring the words which influence each of the categories.

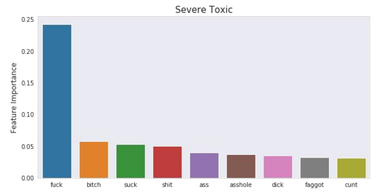
## REFERENCES

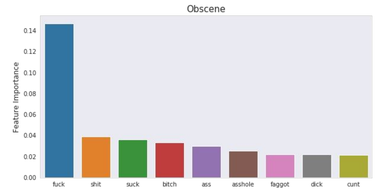
1. "Jigsaw or Conversation AI", Toxic Comment Classification Challenge, https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
2. Stop the S@#$ - Jagan
3. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

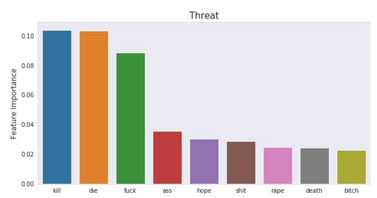
## EXPLORATORY DATA ANALYSIS

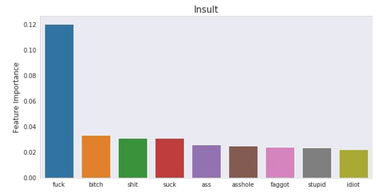
To identify the words which majorly contribute to a category, we performed a feature-importance analysis by calculating the TF-IDF scores for all the words across all the comments, and plotted the top 9 words for these categories.

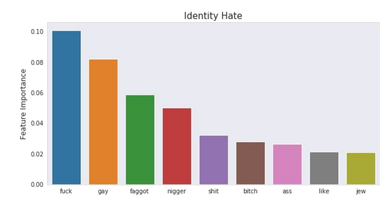


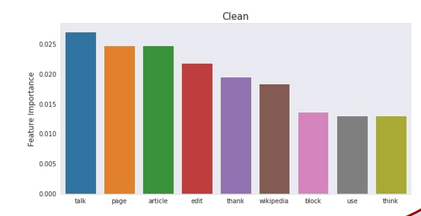



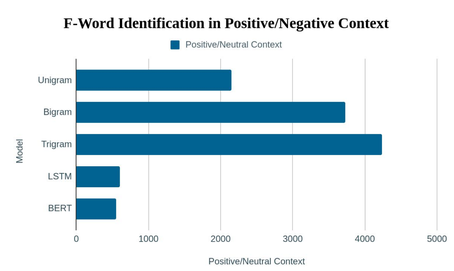

Furthermore, all comments which do not belong to any of these categories, we did a similar analysis to retrieve the "clean" words, and plotted the top-9 clean words as well.




## F-WORD ANALYSIS

Considering the f-word is used in different contexts, both positive and negative, we further checked the performance of the various models in identifying the context of this word. Counts of all the instances in the test-set which didn't belong to any of the categories, and contain the f-word were retrieved, from the predictions made by each of the models.

It was observed that the neural network based models had the least number of such words, indicating that they were able to identify the context of the f-word better than the n-gram models.




## EXPERIMENTATION

The following machine learning models were used for sentiment analysis:

- Naive Bayes
- Long Short Term Memory (LSTM)
- Bidirectional Encoder Representations from Transformers (BERT)

Baseline was developed by using the word lists identified for each category. Word lists with the various lengths were used, and we observed that the maximum such performance was achieved with the Top-10 words, and after 20 words, the performance started decreasing.

Pretrained word2vec embeddings were used with the LSTM model.

| Method | AUC-ROC Score (0-100) |
|---|---|
| Presence of Top 5 words | 60.23 |
| Presence of Top 10 words | 61.78 |
| Presence of Top 20+ words | 60.22 |
| NB using Unigram TF-IDF scores | 88.19 |
| NB using Bigram TF-IDF scores | 84.24 |
| NB using Trigram TF-IDF scores | 82.88 |
| **Neural Network + Word2Vec** | **97.77** |
| **BERT** | **98.24** |

We used the AUC-ROC evaluation metric to assess the extrinsic performance of the language models. This metric tells how much the model is capable of distinguishing between the classes. By analogy, an AUC-ROC score of 0.5 is the worst, i.e, has no class separation whatsoever. This metric is superior to other metrics such as accuracy because it's independent of thresholding.

## FUTURE WORK

Looking at the TF-IDF scores for each class, we note that the most important words are verbs or adjectives. To use this information, we can delete everything except verbs and adjectives from the dataset and re-run the models. As this processed dataset contains lesser noise, it should give better results.