Automatic Filtering of Malicious Comments on the Twitter and Wikipedia Data

Ashish Jain
M.Tech CSE
IIIT DELHI
Okhla Phase III, New Delhi
ashish18052@iiitd.ac.in

Piyush Dhyani
M.Tech CSE
IIIT Delhi
Okhla Phase III, New Delhi
piyush18131@iiitd.ac.in

Sarosh Hasan
M.Tech CSE
IIIT DELHI
Okhla Phase III, New Delhi
sarosh18084@iiitd.ac.in

Abstract

Detection and filtering of abusive language in user-generated online comments have become an issue of importance in recent years. Websites that allow users to leave feedback causes plaguing websites with false comments which harms the online business and overall user experience. Toxic comment classification has become an active research field with many recently proposed approaches. We compare different approaches on comment dataset of Twitter and Wikipedia. We will work on an ensemble based model and word embedding techniques.

1 Introduction

Our area of focus is the study of negative online behaviors, like toxic comments (i.e., comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far We've seen a range of publicly available models. But the current models still make errors, and they don't allow users to select which types of toxicity they're interested in finding (e.g., some platforms may be fine with profanity, but not with other kinds of toxic content). The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user.

2 Dataset

We take two datasets into account to investigate these errors: comments on Wikipedia talk pages presented by Google Jigsaw during Kaggles Toxic Comment Classification Challenge[1] dataset distribution see Figure 2 and a Twitter dataset by

[1]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

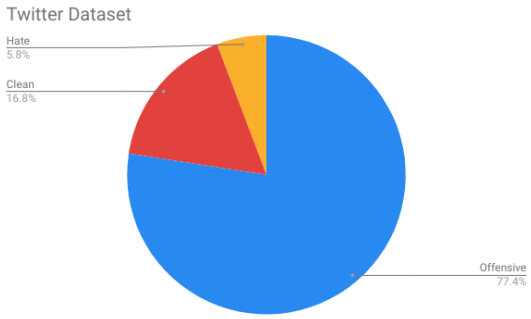

Figure 1: Twitter Data Distribution

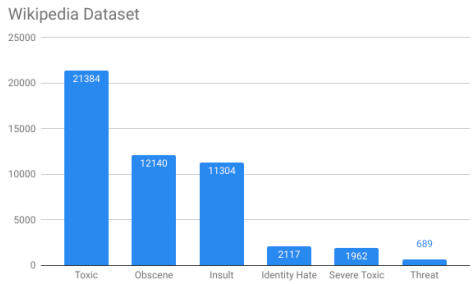

Figure 2: Wikipedia Data Distribution

(Davidson et al., 2017) dataset distribution see Figure 1. These sets include common difficulties in datasets for the task: They are labeled based on different definitions; they include diverse language from user comments and Tweets, and they present a multi-class and a multi-label classification task respectively.

3 Baseline Models

3.1 Using twitter dataset

In the unbalanced twitter dataset the Tf-idf based, Sentimental based, Flesh Reading Ease (FRES),Flesh-Kincaid (F-K) scores are used as features. FRES and FK scores indicates the score for reading ease text. In FRES score high score indicates material is easier to read and less is more

difficult.

$$206.835 + 1.015 * \left(\frac{TotalWords}{TotalSentences}\right) - 84.6 * \left(\frac{TotalSyllables}{TotalWords}\right)$$

FK score is US grade level to check the readability level of various books. If score is greater than ten it indicates no of years of education required to understand the text.

$$-15.59 + 0.39 * \left(\frac{TotalWords}{TotalSentences}\right) + 11.8 * \left(\frac{TotalSyllables}{TotalWords}\right)$$

The number of occurrences of each term along with idf weight (Tf-idf) used as n-gram features with other features. For classification different classifiers were used but the Logistic Regression and Support Vector Machine performs better. Out of which Logistic Regression with L2 Regularization performs best in terms of precision and recall of each class. The results with above features and Logistic Regression is shown in Figure 3 and the confusion matrix is shown in Figure 4.

3.2 Using Multi-label Wikipedia dataset

Multilabel classification problems are the problems in which an instance can assign to various categories, where we have a set of target labels. There are different methods to solve a multi-label classification problem. We use Problem Transformation method in our project. The proposed approach can be carried out in three different ways as:

1. Binary Relevance
2. Classifier Chains
3. Label Powerset

3.2.1 Binary Relevance

This technique treats each label as a separate single class classification problem and classifies the comment as per the regular classifier would.

3.2.2 Classifier Chains

The first classifier is trained just on the input data and then each next classifier is trained on the input space and all the previous classifiers in the chain. The previous class label is then appended as the feature in the comment and then again classified as per the training model.

3.2.3 Label Powerset

In this, we transform the problem into a multi-class problem with one multi-class classifier is trained on all unique label combinations found in the training data. The Problem with this approach is that it is computationally costly.

The result of accuracies using different Multi-Label Classifier on tf-idf based feature are shown in Figure 5.

	Precision	Recall	F1 Score	Support
Hate	0.61	0.37	0.46	164
Offensive	0.91	0.96	0.94	1905
Neither	0.88	0.8	0.84	410
Macro Average	0.8	0.71	0.75	2479
Micro Average	0.89	0.9	0.89	2479

Figure 3: Twitter Data Evaluation Results

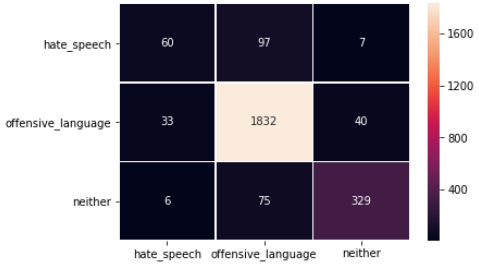

Figure 4: Confusion Matrix Twitter Data

4 Future Work

The work and techniques, which we are planning to execute for getting better results are

4.1 Word Embedding Techniques

Word embedding is a type of mapping that allows words with similar meaning to have similar representation. Here we may try different models with Word2Vec, FastText and GloVe word embedding.

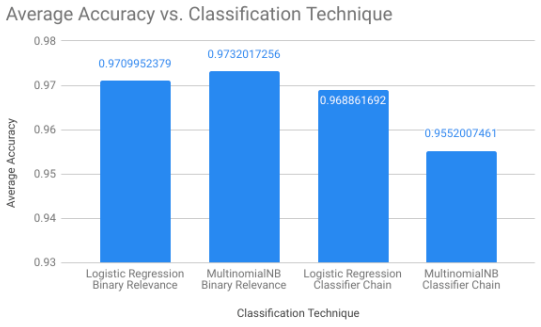

Figure 5: Accuracy of Baseline Model on Wikipedia Data

4.1.1 FastText

FastText is an extension to Word2Vec. Instead of feeding individual words into the Neural Network, FastText breaks words into several n-grams. The word embedding vector will be the sum of all these n-grams. Rare words can now be properly represented as some of their n-grams may appear in other words with a higher chance.

4.2 Ensemble Based Model

Ensemble methodology is to build a predictive model by integrating multiple models. It is well-known that ensemble methods can be used for improving prediction performance in wikipedia multi-label dataset.We will use classifier chain model using different permutations of labels voting.

4.3 In Twitter dataset

We will try to improve the macro Recall and precision of this dataset as to increase the precision and recall of the less frequent class.

References

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.