Detecting toxic contents in online conversations

Amal Sony	Mohd Sharique Khan	Natansh Negi	Siddu Madhure Jayanna
200261394	200261202	200262834	200263301
asony@ncsu.edu	mkhan8@ncsu.edu	nnegi2@ncsu.edu	smadhur@ncsu.edu

1 Problem Statement

An existential problem for any major website today is how to handle toxic and divisive content. Discussing things you care about can be difficult. 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 comments. One 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) and develop a model that's capable of detecting toxicity. We also aim to provide a score of toxicity for these comments if time permits.

2 Related Work

2.1 Deeper Attention to Abusive User Content Moderation

2.1.1 Overview

This research paper aims to apprise us of a way to segregate the abusive content with the non-abusive content by using the concepts of Deep Neural Network. The authors use this to classify their dataset. However, they also mention that using deep neural network is very time consuming as we apply operations on several layers. Additionally, the authors also found that a combination of (RNNLMs, LR with DOC2VEC, NBSVM) delivers better results than using any of these alone.

2.1.2 Task

Different deep neural network approaches discussed in this paper:

- CNN based methods
- RNN based methods
- DETOX implementation
- LIST Baseline

2.1.3 Data

Wikipedia Comments and Gazzetta Comments.

2.2 Challenges for toxic comment classification

2.2.1 Overview

In this research paper the author presented multiple approaches(mentioned in Task) for toxic comment classification. Additionally, the authors also discussed about multi class classification wherein the comments are further segregated according to the severity of toxicity. The approaches make different errors and can be combined into ensemble with improved F-1 measures. The authors have used Glove word embeddings trained on twitter corpus.

2.2.2 Task

Approaches discussed in this paper:

- Shallow Learners
  - Logistic regression
  - RNN based methods
- Deep Learners
  - CNN
  - RNN

2.2.3 Data

Wikipedia Dataset, Twitter dataset.

2.3 Personal Attacks seen at a scale

2.3.1 Overview

In this research paper the author discusses about the below mentioned approaches for performing toxic comment classification. For the LR and MLP models listed below, they simply use bag-of-words representations based on either word or character level n-grams. They also say that simple n-gram features are more powerful than linguistic and syntactic features, hand-engineered lexicons, and word and paragraph embeddings.

2.3.2 Task

Approaches discussed in this paper:

- Logistic Regression(LR)
- Multi-layer perceptrons (MLP)
- Long short term memory recurrent neural networks (LSTM)

2.3.3 Data

Wikipedia Dataset.

3 Approach

For our project which shares the same goal as discussed in the above mentioned papers, we will be moving forward with building different types shallow learners. We will experiment with an ensemble of different shallow learners and methods and also compare their performance on our dataset. We would not be using any stand alone Deep Neural Network classifier as our primary classifier as it is heavy and expensive. Also, we can achieve better results with the ensemble of different shallow classifiers itself. We would be using Doc2Vec to convert tokenized sentences to numerical vectors which would be again be used to train and build our classifiers.

4 Rationale

- Why are we using NLTK to perform data cleaning and tokenization?
  NLTK provides a function called word_tokenize() for splitting strings into tokens. It splits tokens based on white space and punctuation. We also use NLTK functions to remove stop words, handle sentences containing apostrophes and get word counts.
- Why are we using the ensemble of models?
  We can achieve much better results(as shown by the research results mentioned in papers) with the ensemble of different shallow classifiers and these results will be better than the results of stand alone Deep Neural Network classifiers.

- Why are we using Doc2Vec?
  Even though we would be encompassing a large corpus by using embeddings from external source like Glove, we will be generating our own word embeddings using Doc2Vec to get the embeddings more closer and specific to our use case. This also results in increased model learning and our understanding about Word2Vec and Doc2Vec concepts.

5 Dataset

We have picked the Kaggle dataset for the following challenges:

5.1 Toxic Comment Classification Challenge

`https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data`
This data is gathered from Wikipedia comments which have been labeled by human raters for toxic behavior.

Attributes:

- Id: unique comment identifier
- Comment_text: wikipedia comment text
- Toxic: a comment labeled “toxic” has a value 1, otherwise 0
- Severe_toxic : a comment labeled “Severe toxic” has a value 1, otherwise 0
- Obscene : a comment labeled “Obscene” has a value 1, otherwise 0
- Threat : a comment labeled “Threat” has a value 1, otherwise 0
- Insult : a comment labeled “insult” has a value 1, otherwise 0
- indentity_hate : a comment labeled “hate” has a value 1, otherwise 0

To make the dataset consistent, we have merged the last 5 columns, i.e ‘Severe_toxic’, ‘Obscene’, ‘Threat’ , ‘Insult’ , ‘indentity_hate’, into one column ‘toxic’ by doing the OR of all the columns.

Size: 123MB
Train.csv (160k x 8)
Test.csv (160k x 8)

5.2 Quora Insincere Questions Classification

`https://www.kaggle.com/c/quora-insincere-questions-classification/data`
The training data includes the question that was asked, and whether it was identified as insincere (target = 1). The ground-truth labels contain some amount of noise: they are not guaranteed to be perfect.

Attributes:

- qid - unique question identifier
- question_text - Quora question text
- target - a question labeled "toxic" has a value of 1, otherwise 0

Size: 151MB
Train.csv (1.31m x 3)
Test.csv (376k x 2)

6 Hypotheses

Exploring the data revealed that the number of non toxic comments are 10 times more compared to non toxic comments. We are hypothesizing that this might affect the accuracy by introducing bias. Hence we have made the ratio 1:1 for now. We will of-course test this hypothesis by once we are done with the current analysis and start again with the complete dataset.

7 Experimental Design

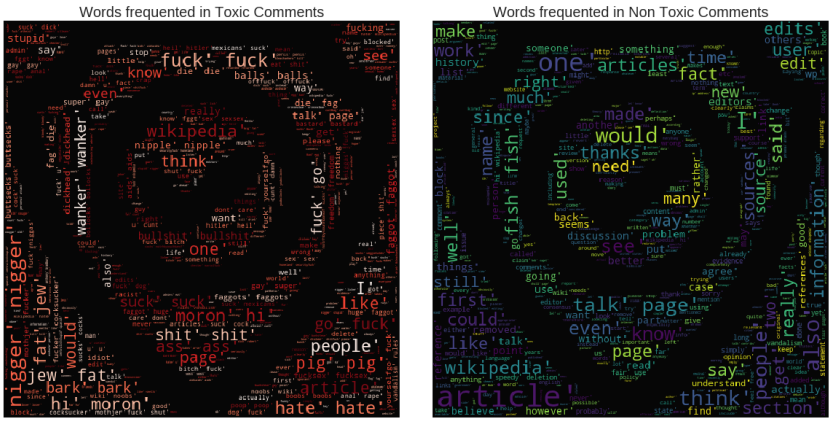
7.1 Steps performed on our dataset

The data set we have is huge, so due to our machine configuration limitations for the time-being we took used about 10% of our dataset. For the dataset 'Toxic Comment Classification Challenge' we have merged the last 5 columns, i.e 'Severe_toxic', 'Obscene', 'Threat' , 'Insult' , 'indentity_hate', into one column 'toxic' by doing the OR of all the columns.

7.2 Data Preprocessing

7.2.1 Data Exploration

As part of this step we checked the presence of null values and also the ratio between the number of toxic and non-toxic comments.



(a) toxic. (b) non toxic.

Figure 1: Word cloud for dataset

7.2.2 Data Cleaning

The data we have is mostly unstructured text, so we have to apply some cleaning on this text using functions of nltk (natural language toolkit). As part of this step we have done the following:

- Converted all of the text to lower-case.
- Tokenized the words in each comment
- Removed the stop words.
- Deleted the punctuations from the text

7.2.3 Data Sampling

In this step, as mentioned above we have performed analysis on a limited sample of the entire dataset (roughly 10%). As we have 2 categories of data we decided to sample data by selecting almost similar number of records of each category (stratified sampling). Additionally we also handled negation and performed stemming for our text.

7.2.4 Data Transformation

As previously mentioned, we have merged the last 5 columns of our dataset, i.e 'Severe_toxic', 'Obscene', 'Threat' , 'Insult' , 'indentity_hate', into one column 'toxic' by doing the OR of all the columns. Additionally we are generating word vectors using DOC2VEC. The DOC2VEC model that we obtain is used in our shallow classifiers.

7.2.5 Model Training

For training our model we have used our pre-processed data and the DOC2VEC model. We have trained our models for the following classifiers:

- Decision Tree Classifiers
- Naive Bayes Classifier
- Logistic Regression
- KNN

7.2.6 Partial Results

The training and test accuracy for the classifiers are:

Classifier	Train Accuracy	Test Accuracy
Decision Tree	100	58.78
Naive Bayes	58.51	52.17
Logistic Regression	80.48	68.93
KNN(K=10)	78.37	54.96

8 Design for Future Experiments

8.1 Dataset

For our future work we would be making use of the complete (training) dataset for training our model. We would also be setting aside validation dataset for tuning hyperparameters of different validation approaches. Testing dataset we have already declared above and we would be using that for testing.

8.2 Hypotheses

Use ensemble methods of shallow classifiers to improve the training and test accuracy.As seen from the results above, the test accuracy lies between 58.78 and 68.93. We hope to improve it by using an ensemble of shallow classifiers. We would be using a combination of different classifiers at each level to check how it affects the overall accuracy.We would also try to segregate the comments based on the level of toxicity, 'Severe_toxic', 'Obscene', 'Threat' , 'Insult' , 'indentity_hate' if time permits.

8.3 Plan of Activities

Amal Sony: Data exploration of the entire dataset and segregate data according to toxicity
Mohd Sharique Khan: explore ensemble methods on shallow classifiers
Siddu Madhure Jayanna: explore ensemble methods on shallow classifiers
Natansh Negi: explore ensemble methods on shallow classifiers

9 References

[1] John Pavlopoulos & Prodromos Malakasiotis & Ion Androutsopoulos (2017) *Deeper attention to abusive user content moderation.* In EMNLP.

[2] Betty van Aken & Julian Risch & Ralf Krestel & Alexander Loser *Challenges for Toxic Comment Classification: An In-Depth Error Analysis.*

[3] Ellery Wulczyn & , Nithum Thain & , Lucas Dixon & *Ex Machina: Personal Attacks Seen at Scale*