Toxicity Comment Classification

By

Seethu Joseph

Contents

Chapter 1

Chapter 2

Chapter 3

Chapter 1

Introduction

1.1 Problem Statement

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. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. 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). Here, I challenged to build a model that's capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate. This model will hopefully help online discussion become more productive and respectful.

1.2 Data

My task is to build classification models which will classify the toxic comments. Given below is a sample of the data set that we are using to classify the toxic comments:

Table 1:1 Toxic comment classification sample data (Columns 1 to 2)

id	**comment_text**
0000997932d777bf	Explanation Why the edits made under my username Hardcore Metallica Fan were reverted? They weren't vandalisms, just closure on some GAs after I voted at New York Dolls FAC. And please don't remove the template from the talk page since I'm retired now.89.205.38.27
000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stuck with. Thanks. (talk) 21:51, January 11, 2016 (UTC)

000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just that this guy is constantly removing relevant information and talking to me through edits instead of my talk page. He seems to care more about the formatting than the actual info.

Table 1:2 Toxic comment classification sample data (Columns 3 to 8)

toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

As you can see in the table below we have the following 8 variables, using which we have to correctly classify the toxic comments:

Table 1:3 Predictor Variables

1. id
2. comment_text
3. toxic
4. severe_toxic
5. obscene
6. threat
7. insult
8. identity_hate

The details of data attributes in the dataset are as follows:

Table 1:4 Data used in the study

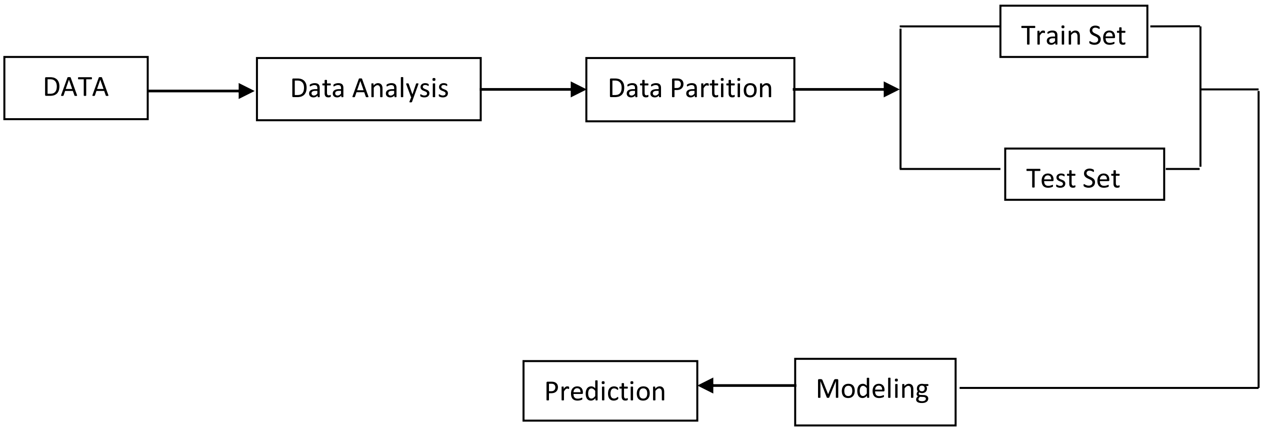
Variable	**Type**	**Description**
Id	Numeric	Id of the each comments
comment_text	Categorical	Wikipedia comments.
Toxic, severe_toxic, obscene, threat,	Numeric	toxicity of each comment_text

insult, identity hate(other 6 variables)		

Chapter 2

Methodology

Figure 1: Predictive Model



2.1. Missing value Analysis

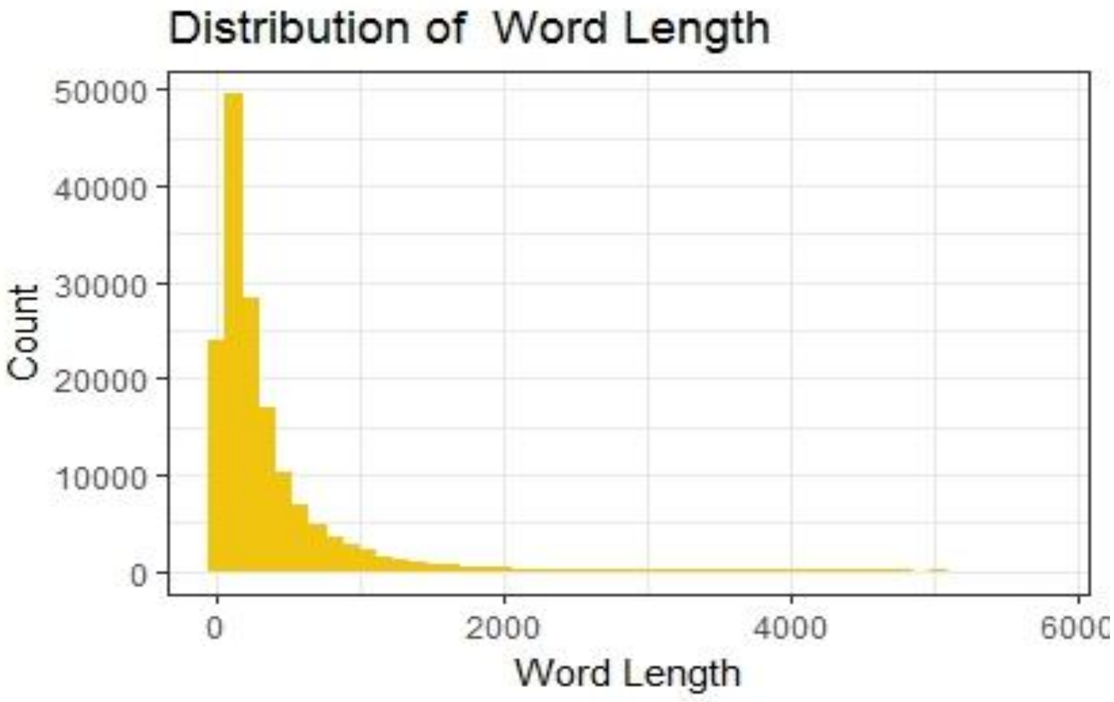
Missing values occur when no data value is stored for the variable in an observation. Missing values are a common occurrence, and you need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Typically, ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values. I could not find missing value in this dataset.

2.2 Exploratory Data Analysis

First step in any analytical project is to understand data using exploratory data analysis. Once you understand data then you can drive data according to the problem statement. Exploratory data analysis is an approach for summarizing and visualizing the important characteristics of a data set. Exploratory data analysis focuses on exploring data to understand the data's underlying structure and variables, to develop intuition about the data set, to consider how that data set came into existence, and to decide how it can be investigated with more formal statistical methods. To accomplish this task (Understand the data), we have graphical and non-graphical tools. Graphical tools includes all visualizations like scatter plot for correlation, boxplot for outlier analysis, histogram for distribution, bar plot for missing value analysis, heat map for locations. Non-Graphical tools include all central statistics and statistical techniques/test.

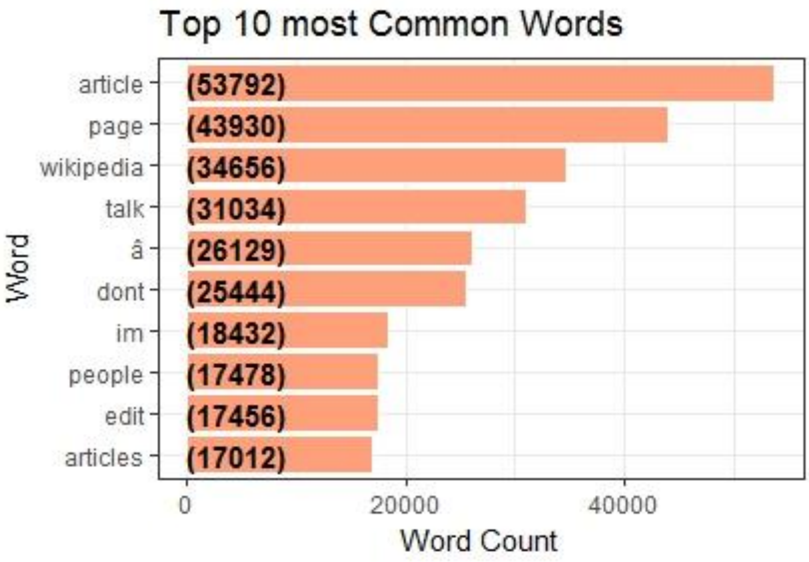
The below figure shows the sentence length distribution of comment text.

Figure 2: Sentence length distribution



Plotting the top ten most common words using ggplot function will give a clear idea about the commonly used words. The below figure shows the most commonly used 10 words in the given data set of toxic comment classification.

Figure 3: Top ten most common words



The above figure shows that some stop words are indicated in the plotting, so we have to eliminate that words. Otherwise that will affect our modelling so I am moving to text mining.

2.3 Text Mining

Text mining is a process of extracting interesting and non-trivial information and knowledge from unstructured text and also identifying novel information from a collection of texts (also known as a corpus). It discovers useful and previously unknown "gems" of information in large text collections. Unlike the data mining, pre-processing techniques are different for text. First we need to convert text (unstructured data) into structured (Rows and Columns) using pre processing steps then only we can able to apply machine learning algorithms.

Text Preprocessing Steps**:**

1. Remove Numbers
2. Remove Punctuation Marks
3. Case Folding (Converting uppercase letters to lower case)
4. Remove stop words
5. Stemming
6. Lemmatization
7. Strip white space
8. Synonym check
9. Remove metadata

Once you perform above steps then convert cleaned text into document term matrix (data frame/Matrix or structured data) using simple term frequency method or weighting method.

Toenization of the sentence means splitting sentence into words and that will give the information about the frequency of the each words in particular variable of given dataset.

Similarly we can group the unique categories of the word. The combination of the toxic, severe toxic, insult, obscene, threat, identity hate will create unique categories.

I wish to find out the important words in this Toxic Comments. For Example, for your young child, the most important word is mom. Example for a bar tender, important words would be related to drinks. Explore this using a fascinating concept known as Term Frequency - Inverse Document Frequency. A document in this case is the set of lines associated with a unique category determined by the various elements such as toxic, severe toxic, obscene, threat, insult and identity hate.

TF-IDF computes a weight which represents the importance of a term inside a document. It does this by comparing the frequency of usage inside an individual document as opposed to the entire data set (a collection of documents). The importance increases proportionally to the number of times a word appears in the individual document itself–this is called Term Frequency. However, if multiple documents contain the same word many times then you run into a problem. That's why TF-IDF also offsets this value by the frequency of the term in the entire document set, a value called Inverse Document Frequency.

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document)
IDF(t) = log_e(Total number of documents / Number of documents with term t in it).
Value = TF * IDF

Found the twenty five most important words using TF-IDF. The below figure shows the plotting.

Figure 4: Twenty five most important words

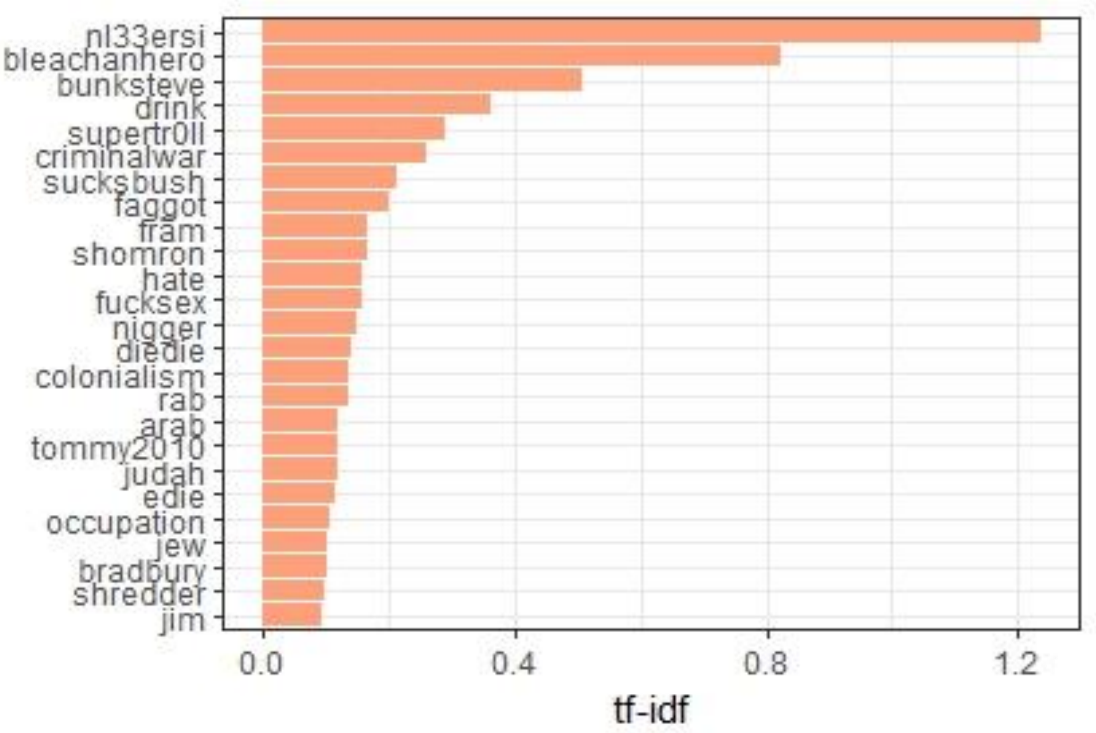
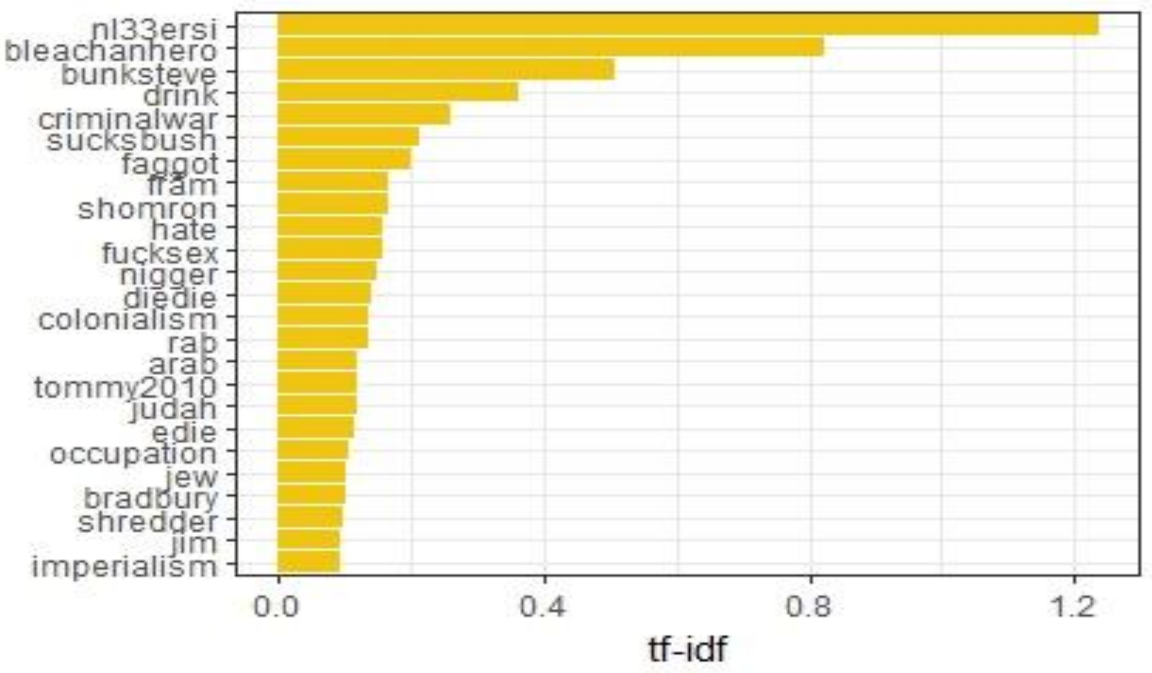
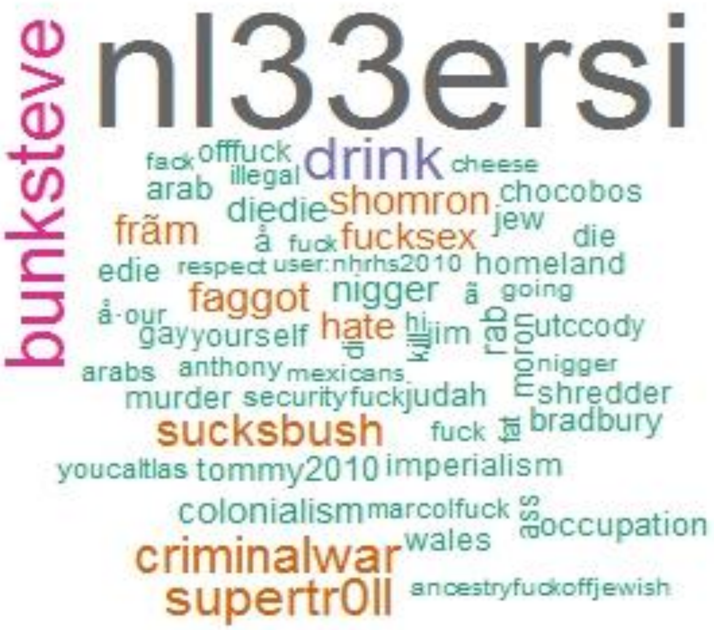


Figure 5: TF-IDF of the toxic comments



The below wordcloud shows the 60 most important terms based on TF-IDF score. Higher the score bigger will be the size of the text

Figure 6: Wordcloud of the most important words



Create the document term matrix for train and test data. A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. There are various schemes for determining the value that each entry in the matrix should take. One such scheme is TF-IDF and they are useful in the field of natural language processing.

2.4 Modelling

Modelling of the data is done by using decision tree method. Trained the data using this method and classify the toxicity of comment. The training data is prepared by above mentioned methods.

2.4.1 Decision Tree

A decision tree is a graphical representation of possible solutions to a decision based on certain conditions. It's called a decision tree because it starts with a single box (or root), which then branches off into a number of solutions, just like a tree. It is a predictive model based on a branching series of Boolean tests. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Each internal node (decision nodes) represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test and each leaf nodes represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Chapter 3

Conclusions

Both industrial and research community in the last few years have made several tries to identify an efficient model for online toxic comment prediction due to its importance in online interactive communications among users, but these steps are still in their infancy. This work is devoted to the study of a recent approach for text classification involving word representations and decision tree method. I investigate its performance in comparison to more widespread text mining methodologies for the task of toxic comment classification. This methodology can provide enough evidence that it is appropriate for toxic comment classification. The results are motivating for further development of technologies for mining in the near future.

3.1 Model Evaluation

There are several criteria that exist for evaluating and comparing models. In our industry, we consider different kinds of metrics to evaluate our models. The choice of metric completely depends on the type of model and the implementation plan of the model. According to our problem statement, it is a classification and we have to choose confusion matrix for evaluation.

3.1.1 Confusion Matrix

It is also called error matrix or transition matrix. It helps to evaluate the performance of the classification model. It shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data.

According to our application, accuracy is the main factor evaluating the model performance.

3.1.1.2 Accuracy

The proportion of the total number of predictions that was correct. It say How accurately model can able to classify.

3.2 Model Selection

While analysing the model, I got accuracy of decision tree is 82% ,so I opted this model for classifying the toxic comments.

References

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