# **SOCIAL MEDIA COMMENTS**

# **CLASSIFICATION PROJECT**

# AIDI 1002

# STATEMENT OF WORK

**Product Manager: Marcos Bittencourt**

**Team Member(s): Esra Ersoy (100812358)**

**PROBLEM STATEMENT**

Social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Although it is commonly known that hate is a problem across multiple platforms, there is a lack of models for online hate detection.  
Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behaviour.  
There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.  
While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it.

**PROJECT OBJECTIVE**

The problem I am planning to solve is the tagging of internet comments that are aggressive towards other users. My goal is to build a prototype of online hate and abuse comment classifier which can be used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

Goals:

1. Explore the dataset to get a better picture of how the labels are distributed, how they correlate with each other and what defines negative or non-negative comments.

2. Create a baseline score with a simple logistic regression classifier.

3. Explore the effectiveness of multiple machine learning approaches and select the best for this problem.

4. Select the best model and tune the parameters to maximize performance.

5. Build the final model with the best performing algorithm and parameters and test it on a holdout subset of the data.

**DATASET REQUIREMENTS AND DESCRIPTIONS**

Data requirement for this project is training and test datasets consist of labelled social media comments in order to work on a classifier model. I chose my dataset from Kaggle.

The data set I chose contains 2 files as training set and test set. Training set has approximately 159,000 records.

Records contain 8 fields which includes ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’.  
The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment.  
The data set includes:

Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.

Highly Malignant: It denotes comments that are highly malignant and hurtful.

Rude: It denotes comments that are very rude and offensive.

Threat: It contains indication of the comments that are giving any threat to someone.

Abuse: It is for comments that are abusive in nature.

Loathe: It describes the comments which are hateful and loathing in nature.

ID: It includes unique Ids associated with each comment text given.

Comment text: This column contains the comments extracted from various social media platforms.

As the given test set has no record labels in it, I am planning to split the training data set as training, validation and test sets.

Data Source: <https://www.kaggle.com/surekharamireddy/malignant-comment-classification?select=train.csv>

**DATA LIMITATIONS AND CONSTRAINTS**

**Data Errors:** When we examine the dataset, it is clearly seen that it has some errors. Despite the fact that some comment texts having only punctuations and does not include any words inside, they were labeled as one of the malignant labels. Therefore, data cleansing is essential for this dataset, otherwise the accuracy level will not be high.

**Limited Number of Negative Labelled Records:** Compared to entire dataset the number of records labelled as ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’ are relatively low. Just over 10% of this dataset is labeled as negative, but some of the subcategories are extremely rare making up less than 1% of the data. Because of this imbalance, accuracy is a practically useless metric for evaluating classifiers for this problem.

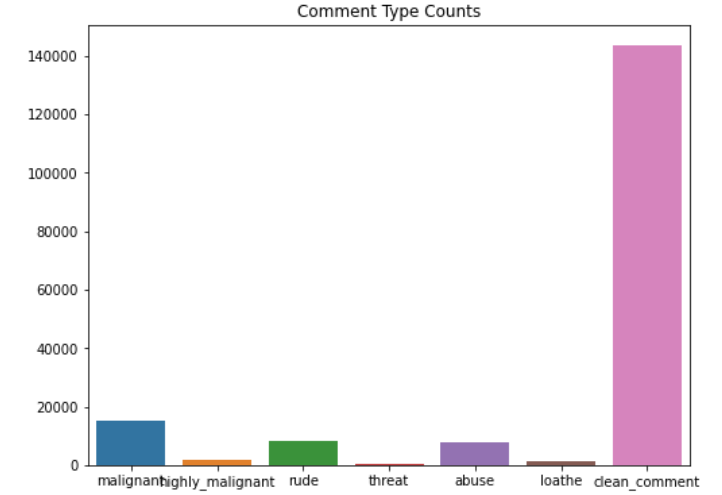
**EXPLORATORY DATA ANALYSIS**

Data Exploration on this dataset contains 159,571 comments from various social media sources.

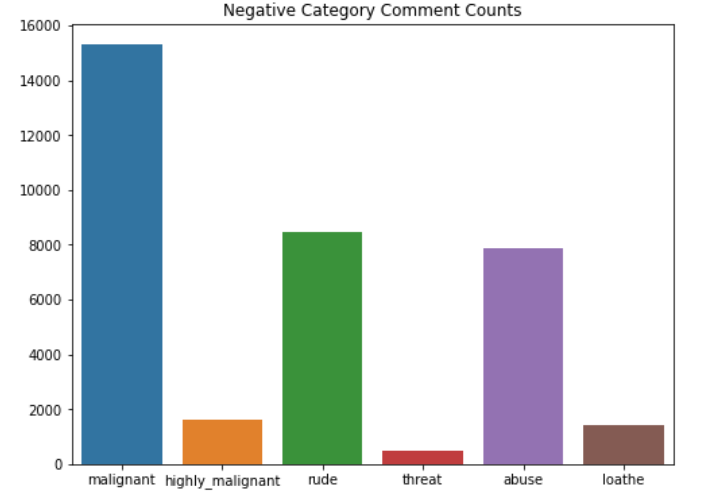
The data consists of one input feature, the string data for the comments, and six labels for different categories of toxic comments: Malignant, highly malignant, rude, threat, abuse and loathe. The figure below contains a breakdown of how the labels are distributed throughout the dataset, including overlapping data. While most comments with other labels are also malignant, not all of them are. Only “highly malignant” is clearly a subcategory of “malignant.” And it’s not close enough to be a labeling error. This suggests that “malignant” is not a catch-all label, but rather a subcategory in itself with a large amount of overlap.

I created a seventh label called “clean\_comment” to represent the comments which have no negative labels on it. From here on in, I’m going to refer to any labeled comments as negative.

Label distribution between negative labels and ‘Clean’ label:

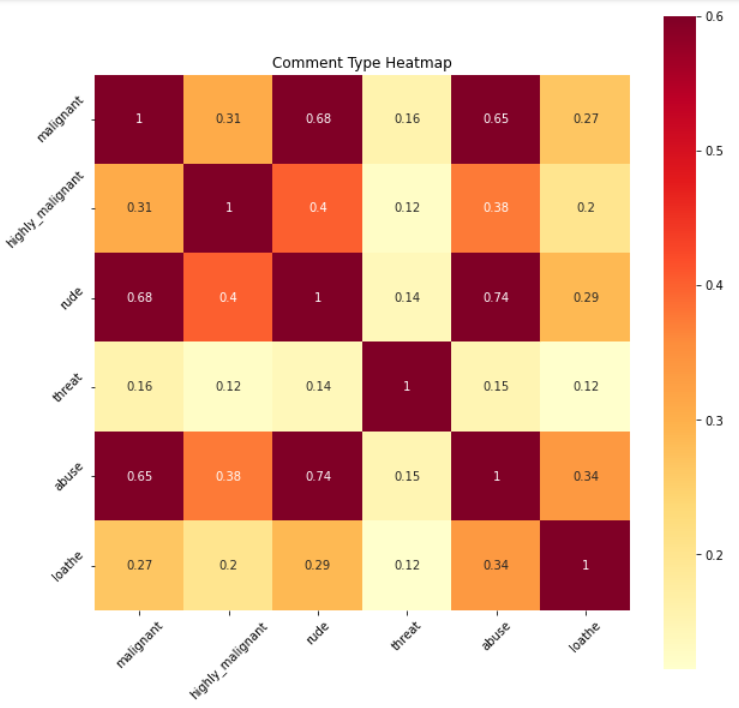


Label distribution only between negative labels:



The correlation matrix below provides more insight into these overlapping categories. ”Threats” are not likely to be “higly malignant”, nor are they likely to be “rude” or “loathe”. But “abuse” comments are often “rude”, and “rude” comments are usually “malignant”. “threat” labels doesn’t really have much overlap at all.

I believe the categories with significant overlap will be more difficult to predict, as they’ll have similar contributing features, but “threat” and “loathe” labels will have more unique attributes and be easier to predict.

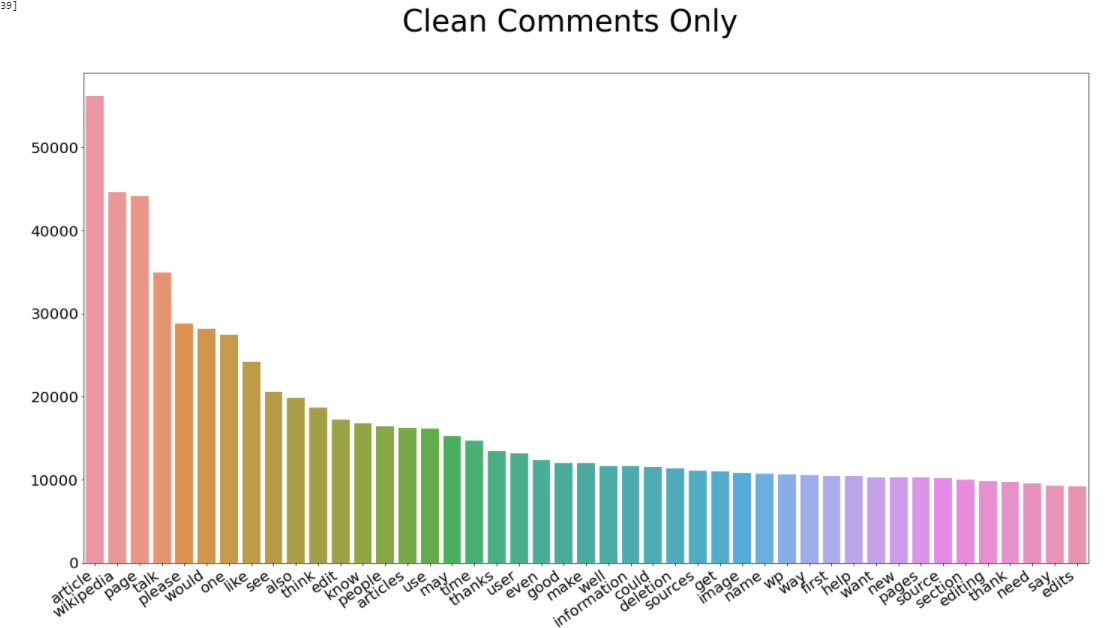


**NLP PREPROCESSING**

**Feature Extraction**

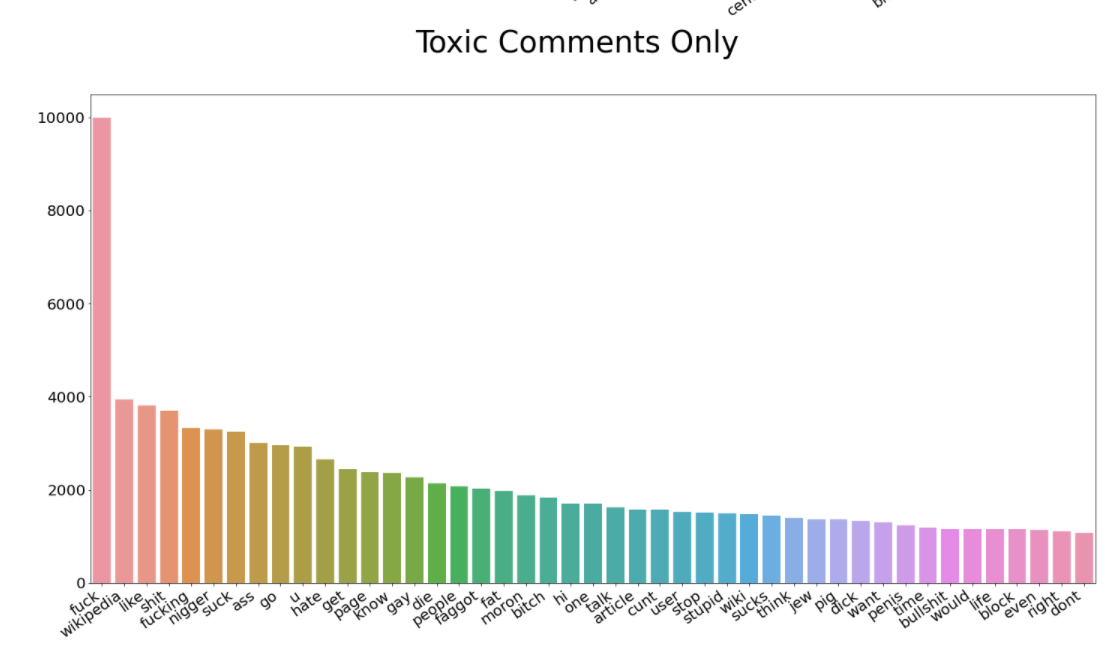
Before doing further analysis, we need to transform text data into numerical data. Machine learning algorithms operate on a numeric feature space, expecting input as a two-dimensional array where rows are instances and columns are features. In order to perform machine learning on text, we need to transform our text comments into vector representations such that we can apply numeric machine learning. More simply, vectorization, and is an essential first step toward language-aware analysis.

After elimination of stopwords, word vectorization provided us a clear number of words recurring in each text comment.



Article, page, please, think, edit, etc. The highest frequency words are about what you would expect from people discussing in social media platforms.

Now let’s look at what a bad comment looks like. The difference in the highest frequency vocabulary is stark.



**Feature Engineering for Text Classification and NLP**

Feature Engineering methods enables us to reach higher accuracy rates with our model. Within a text, punctuations, lower/upper cases, stop words which doesn’t have any meaning or value to enrich our model have to be removed from the data. We can also do some other feature engineering like, counting the number of emojis used, type of emojis used, what frequencies of unique words, etc. We can define our features by analyzing the dataset.