

MALIGNANT COMMENTS CLASSIFIER

Submitted by:

Anusha K

**INTRODUCTION**

* Business Problem Framing

Our goal is to build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

* Conceptual Background of the Domain Problem

The proliferation of 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 researchers have found 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.

Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as unoffensive, but “u are an idiot” is clearly offensive.

**Analytical Problem Framing**

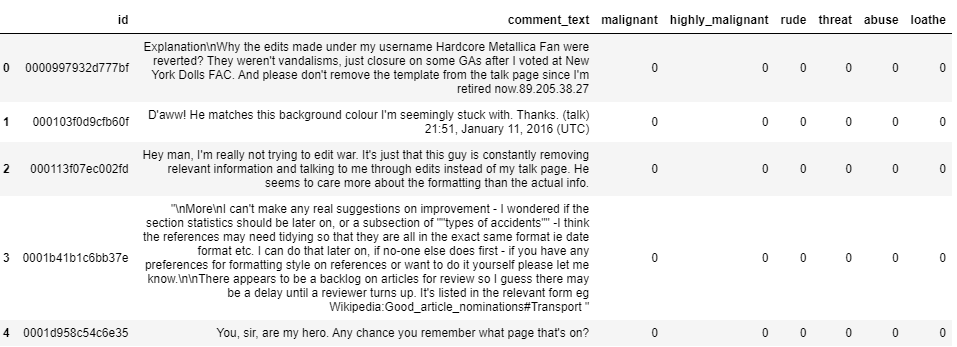
* Mathematical/ Analytical Modeling of the Problem

This is a multilabel Classification Problem, there are 6 target variables with 0 and 1 labels in each of the target variable. The input data is of text type and there is only one input attribute that is comments extracted from various social media platforms.

* Data Sources and their formats

The input is only the text data and 6 target variables. This is a multilabel classification problem. Target variables are of int type which consists only 0 and 1 values. Input variable is of Text format.

The data look like below:



* Data Preprocessing Done

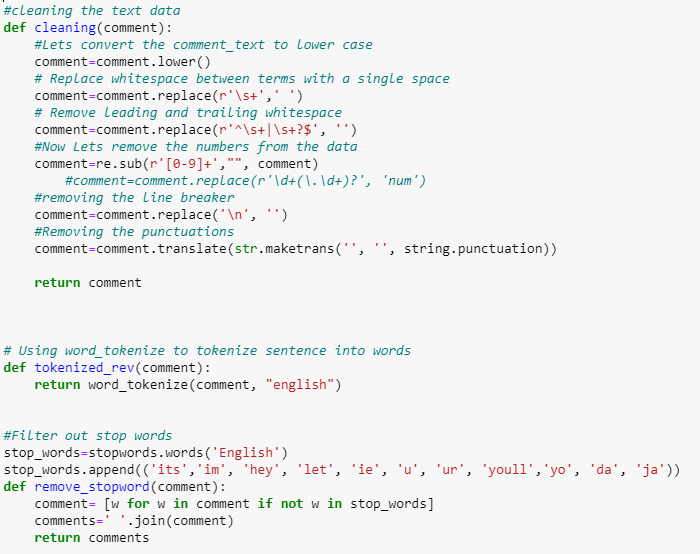
As the input is text, we need to perform some of the data cleaning and pre-processing steps to get the clean data to feed it to the model.

The target values are clean, we need not perform any cleaning or pre-processing.

As part of the data cleaning and pre-processing, initially the input text data must be converted to lowercase.

The test data when converted to vectors there is no meaning for the punctuations, hence the punctuations are removed from the text. The whitespaces between the terms are replaced with a single space. The Numbers does not have any meaning for this problem and removed from the data. Line breaker also has been removed from the data.

Once the data is cleaned, the text data is passed through the tokenizer. Here we have used word\_tokenizer to tokenize the sentences into words. From this list of words, stopwords has been removed.



Now we have the clean data.

* Data Inputs- Logic- Output Relationships

The login behind this problem is that, if a comment has more hatred, rude or abusive kind words the target variables turn to 1. If the comment is in good form and does not have any negative impact, then the target variables turns to 0.

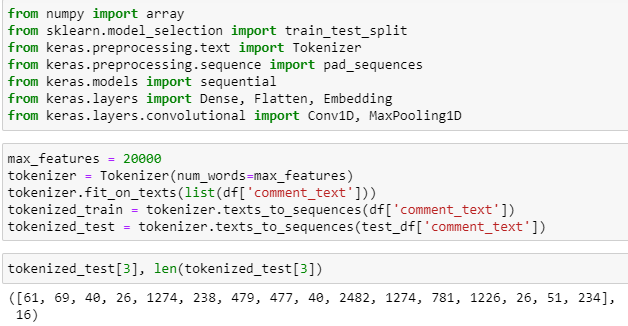
**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

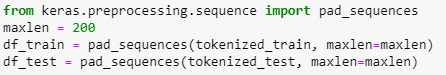
The dataset has 6 target variables, and each target variable has 2 classification 0 and 1, hence this is a multi-label classification problem. As we have text input, the data has been cleaned and then building the model and selecting the best model.

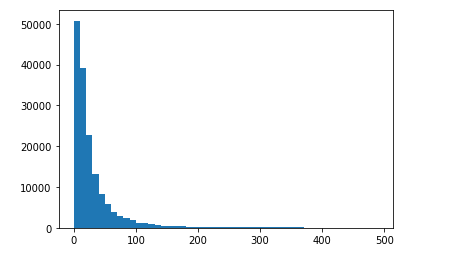
* Testing of Identified Approaches (Algorithms)

In this problem we have used deep learning Sequential algorithm as we have 6 target variables to be predicted. We have used Tokenizer from Keras text Preprocessing library to convert the text to vector form.



Some of the comments are too long and some are too short. But we need to feed a stream of data that has a consistent length i.e., fixed number of features. So, we use ‘Padding’. We could make the shorter sentences as long as others by filling the shortfall by zeroes. On the other hand, we must trim the longest sentences to a fixed size length called maxlen.



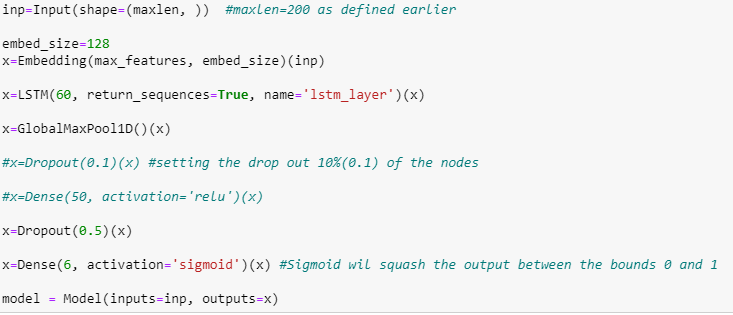


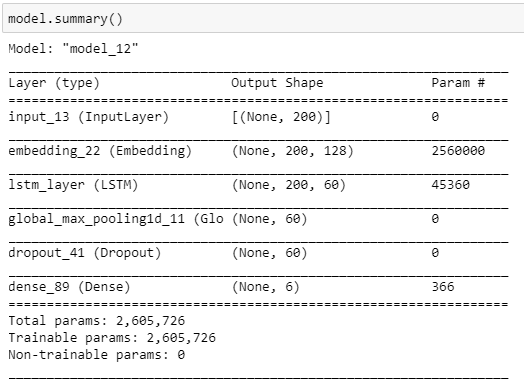
The graph above shows the word length in each comment. The maxlen greater than 200 are rare occurrences. Hence, we can use 200 as maxlen.

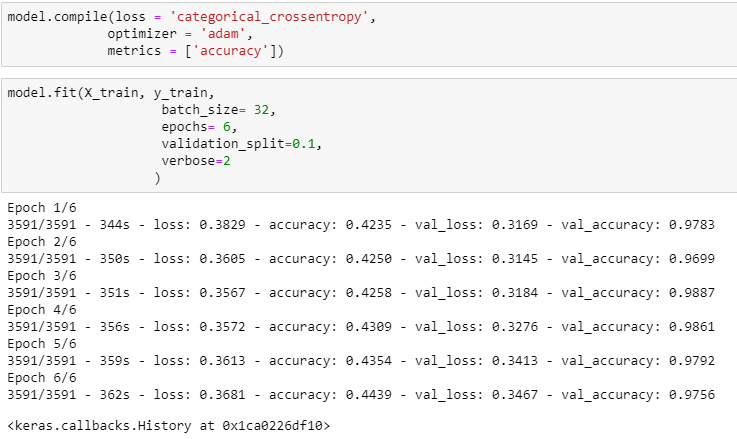
Then the dataset was split into training set and test set using train\_test\_split method.



Now the training set is trained using the deep learning Sequential model.

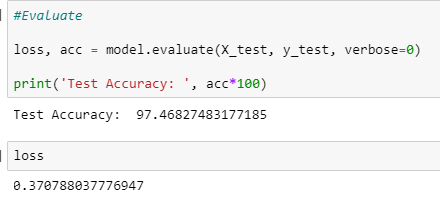






* Evaluation:

By using this model, we have achieved the test accuracy score of 97.468% with a loss of 0.3708%.



**CONCLUSION**

Deep learning models are commonly used for the multi-labelled classification problems. This helps to achieve the best accuracy score with minimal loss.