MALIGNANT COMMENT CLASSIFIER PROJECT

Prepared by Pratyush Raj

Data Science Intern at Flip Robo Technologies

SME Name: Mr, Kashif

Acknowledgment

It is my deepest pleasure and gratification to present this report. Working on this project was an incredible experience that has given me very informative knowledge regarding the data analysis process.
All the required information and dataset are provided by **Flip Robo Technologies** which helped me to complete the project. I want to thank my SME
for giving the dataset and instructions to perform the complete case study process.

INTRODUCTION

Problem Statement:

The background for the problem originates from the multitude of online forums, where-in people participate actively and make comments. As the comments sometimes may be abusive, insulting, or even hate-based, it becomes the responsibility of the hosting organizations to ensure that these conversations are not of a negative type. The task was thus to build a model which could make predictions to classify the comments into various categories. Consider the following examples:

The exact problem statement was thus as below:

Given a group of sentences or paragraphs, used as a comment by a user in an online platform, classify it to belong to one or more of the following categories — toxic, severe-toxic, obscene, threat, insult, or identity-hate with either approximate probabilities or discrete values (0/1).

Introduction:

Online forums and social media platforms have provided individuals with the means to put forward their thoughts and freely express their opinion on various issues and incidents. In some cases, these online comments contain explicit language which may hurt the readers.
Comments containing explicit language can be classified into myriad categories such as Toxic, Severe Toxic, Obscene, Threat, Insult, and Identity Hate. The threat of abuse and harassment means that many people stop expressing themselves and give up on seeking different opinions.
To protect users from being exposed to offensive language on online forums or social media sites, companies have started flagging comments and blocking users who are found guilty of using unpleasant language. Several Machine Learning models have been developed and deployed to filter out the unruly language and protect internet users from becoming victims of online harassment and cyberbullying.

Objective:

The main purpose of building this model is to prevent abusive comments which in turn will Detroit the mindset of an individual or people, nowadays a lot of abusive and lethargic comments can be seen on various social media platforms which creates a negative environment among the people and community, so to stop this type of activity a machine learning model is built to identify the malignant text and filter it out as soon as it encounters it.

Review of Literature:

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 Using Machine Learning Techniques. In this we are investigating the application of supervised machine learning techniques to predict the comments. The predictions are based on historical data collected from websites like twitter etc. Different techniques. To build a model for predicting the comments we have used Supervised machine learning.

Motivation:

Online platforms when used by normal people can only be comfortably used by them only when they feel that they can express themselves freely and without any reluctance. If they come across any kind of a malignant or toxic type of a reply which can also be a threat or an insult or any kind of harassment which makes them uncomfortable, they might defer to use the social media platform in future. Thus, it becomes extremely essential for any organization or community to have an automated system which can efficiently identify and keep a track of all such comments and thus take any respective action for it, such as reporting or blocking the same to prevent any such kind of issues in the future.

Description of dataset

ws and 2 columns for testing. I observed that every 1 in 10 samples was toxic, every 1 in 50 samples was obscene and insulting, but the occurren ces of sample being severe-toxic, threat and identity hate was extremely rare.

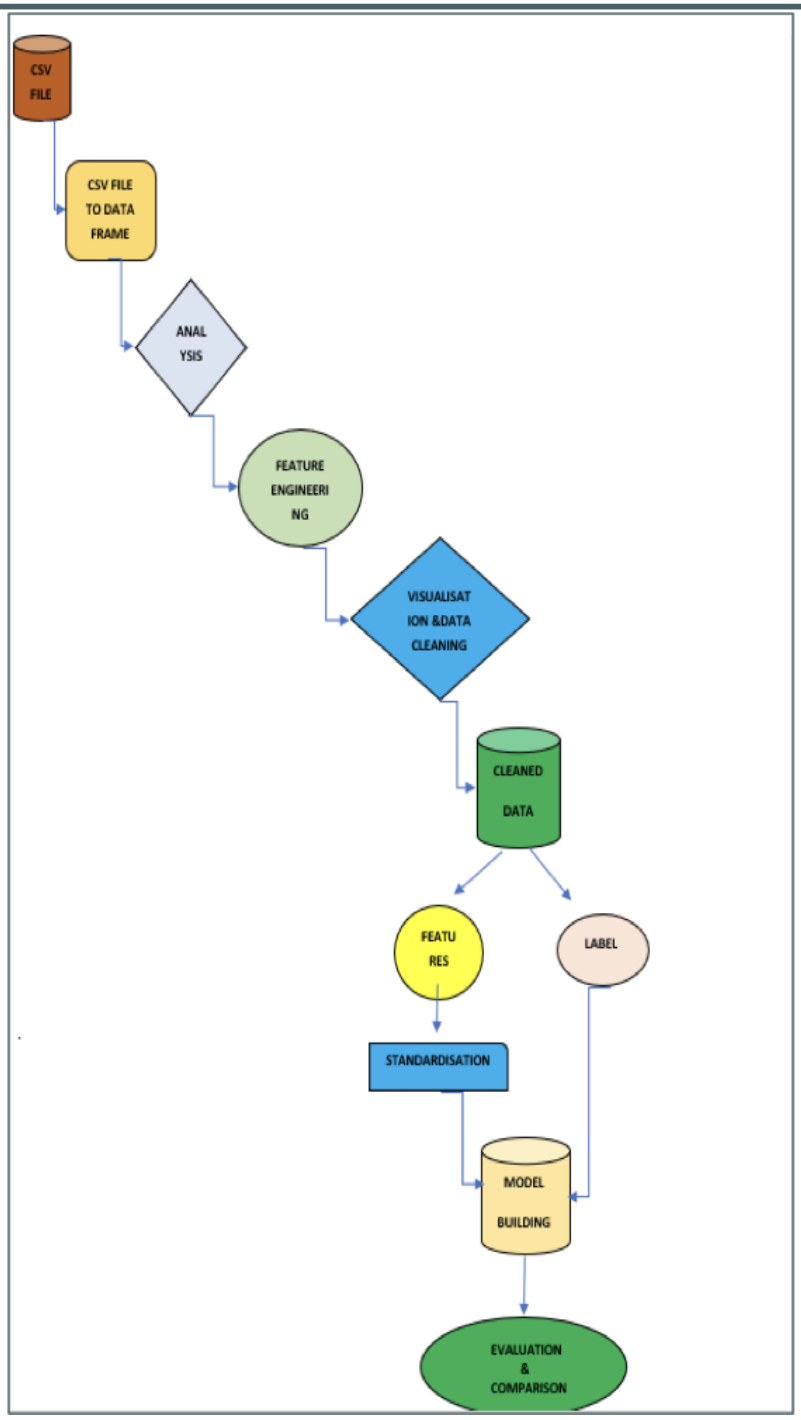
The dataset consists of the following fields-

- **id: An 8-digit integer value, to get the identity of the person who had written this**

  comment

- **comment text: A multi-line text field which contains the unfilte red comment.**
- **malignant: binary label which contains 0/1 (0 for no and 1 for y es).**
- **highly malignant: binary label which contains 0/1.**
- **rude: binary label which contains 0/1.**
- **threat: binary label which contains 0/1.**
- **abuse: binary label which contains 0/1.**
- **loathe: binary label which contains 0/1**

Out of these fields, the comment text field will be pre-processed and fitted into different classifiers to predict whether it belongs to one or more of the labels/outcome variables (i.e., malignant, highly mal ignant, loathe, threat, abuse and rude). We have a total of 15957 1 samples of comments and labelled data, which can be loaded from train.csv file. The first 5 samples are as follows.


CSV FILE
CSV FILE TO DATA FRAME
ANAL YSIS
FEATURE ENGINEERI NG
VISUALISAT ION &DATA CLEANING
CLEANED DATA
FEATU RES
LABEL
STANDARDISATION
MODEL BUILDING
EVALUATION & COMPARISON

CONCLUSION:

Communication is one of the basic necessities of everyone's life. People need to talk and interact with one another to express what they think. Over the years, social media and social networking have been increasing exponentially due to an upsurge (rise) in the use of the internet. Flood of information arises from online conversation on a daily basis, as people are able to discuss, express themselves and express their opinion via these platforms. While this situation is highly productive and could contribute significantly to the quality of human life, it could also be destructive and enormously dangerous. The responsibility lies on the social media administration, or the host of organization to control and monitor these comments.

This research work focuses on developing a model that would automatically classify a comment as either malignant or non-malignant using logistic regression. Therefore, this study aims to develop a multi-headed model to detect different types of malignant comment like threats, rude, abusive, and loathe. By collecting and preprocessing malignant comments for training and testing using term frequency- inverse document frequency (TF-IDF) algorithm, developing a multi-headed model will detect different types of malignant comment using logistic regression to train the dataset, and evaluate the model using confusion metrics

LIMITATION:

As per my understanding and study of dataset I presented the best model with