FLIP ROBO

Malignant Comments Classifier

Submitted by:

Nirav Mehta

ACKNOWLEDGMENT

First of all, I would want to thank my data trainer and SME Khushboo Garg ma’am for giving me this opportunity. Their ongoing direction and counsel were extremely helpful in making the project's implementation. I received helpful advice from Data Trained that was essential to creating the most error-free version of my report. The live Data skilled assistance that helped me with this project and showed me how to handle the skewness deserves my gratitude as well.

INTRODUCTION

Although a sizable number of internet comments found in public spaces are often positive, a sizable portion are toxic in nature. Datasets are gathered from flip flops and then processed to eliminate noise. The term frequency-inverse document frequency (TF-IDF) technique, a machine learning method, is used to train the dataset by processing the dataset in the form of transformation of raw comments before feeding it to the Classification models. The comments contain a lot of errors, which increases the number of features by a factor of two. The processed dataset is trained using the logistic regression technique to distinguish between malicious and benign remarks.

People now have the ability to openly voice their opinions on a variety of problems and events through online forums and social media platforms. These internet comments occasionally use explicit language that could offend readers. There are other categories that comments using explicit language can fall under, including Toxic, Severe Toxic, Obscene, Threat, Insult, and Identity Hate. Due to the fear of harassment and abuse, many people refrain from speaking their minds and give up on exploring other points of view.

Companies have started flagging remarks and barring individuals who are found guilty of using foul language in order to prevent users from being exposed to inappropriate language on internet forums or social media platforms. To filter out the foul language and shield internet users from experiencing online harassment and cyberbullying, several Machine Learning models have been created and put into use.

In order to stop this kind of activity, a machine learning model was developed to recognise malicious text and filter it out as soon as it is encountered. The main goal of building this model is to prevent abusive comments that will, in turn, Detroit the mind-set of an individual or people. Today, a lot of abusive and lazy comments can be seen on various social media platforms, which create a negative environment among the people and community.

Our objective is to create a prototype of an online hate and abuse comment classifier that can be used to categorise hateful and offensive comments in order to limit and restrict the spread of hatred and cyberbullying. In this study, we explore the use of supervised machine learning methods to anticipate comments. The forecasts are based on past information gathered from sites like Twitter and others. Several methods we used supervised machine learning to create a model for predicting the comments.

Analytical Problem Framing

Online platforms can only be used by regular people in a way that makes them feel comfortable expressing themselves openly and without hesitation. They might decide not to use the social media platform again if they encounter any kind of poisonous or malicious comment, which could also be a threat, insult, or other form of harassment. Therefore, having an automated system that can effectively recognise and keep track of all such remarks and take appropriate action for it, such as reporting or blocking the same, to prevent any such concerns in the future, becomes imperative for any company or community.

For training, I used a dataset with 159571 rows and 8 columns, and for testing, I used 153164 rows and 2 columns. I found that toxic samples were 1 in 10 and obscene and offensive samples were 1 in 50, but severe toxic, threatening, and identity-hating samples were incredibly rare.

The fields in the dataset are as follows:

- **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 unfiltered comment.
- **Malignant**: binary label which contains 0/1 (0 for no and 1 for yes).
- **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

From these fields, the comment text field is pre-processed and fitted to various classifiers into one or more label/outcome variables (e.g., malicious, highly malicious, disgusting, threatening, abusive, rude, etc.). . There are a total of 159571 examples of commented and labelled data that can be loaded from the train.csv file.

Data processing and feature engineering are important in machine learning to create predictive models. Also, you can't create a model without doing some data processing. For example, we failed to train the model before handling missing values and converting the text in the dataset to numbers, as shown in our experiments. Therefore, our experiments show that pre-processing the data improves the prediction accuracy.

Info: Used to provide information about non-zero values and feature data types. This is the processed comment (object) data type, the other data type is int.

Descriptive methods are used to calculate statistical data such as percentiles, means, and standards of numbers in a series or data frame. Analyse both numeric and object series, and also data frame column sets with mixed data types. There is also information about data distribution.

Converts all uppercase text in comments to lowercase. This is an important technique because it helps clean the data. Recently, one of my readers noticed that depending on the case of the input (e.g. "Canada" vs. "Canada"), she would get different kinds of output, or none at all. This occurred because the dataset contained the case-sensitive word "Canada" and there was insufficient evidence that the neural network could effectively learn less common versions of the weights. There is a possibility. This kind of problem always happens when the dataset is fairly small, and lower case is a great way to deal with sparsity issues.

Now that we have verified that there are no missing records in the data, we decide to start pre-processing the data. First, because comments from online forums usually contain inconsistent language, the use of special characters (like @argument) instead of letters, and the use of numbers (like .n0t) to represent letters. I decided to normalize the text data. I decided to use regular expressions to fix such data discrepancies.

A stop word is a word commonly used in both written and oral communication that does not have a positive or negative effect on what we are saying. A set of words used. Examples of English stop words are "a", "the", "is", "are", etc. The intuition behind using stop words is to remove words with little information from the text and instead focus on the important words. Words can concentrate. For example, if the query is in the context of a search engine, "What is word processing? This can be achieved by not parsing all words in the stop word list. Stop words are widely used in search systems, text classification applications, topic modelling, topic extraction, and more.

Surface lemmatization is very similar to stemming, the goal is to remove inflections and map words to their root forms. The only difference is that lemmatization tries to do it the right way. It doesn't just cut things, it actually converts the words into their actual roots. For example, the word "better" maps to "better." Dictionary such as B. Use word mesh for mapping or some special rule-based approach.

Above is an architecture showing the key steps involved in data pre-processing and model building. When I first analysed the dataset, I found that the data type for a particular column was incorrect, so I fixed it and used feature engineering to create a numeric column from the object column. This is because machine learning models cannot understand text. Understanding You have used charts to visualize and understand your data. Data cleansing occurs after assignment. Data cleansing is done by detecting outliers, checking for skewness, and removing unnecessary columns, but model accuracy is worth nothing. After cleaning the data, we split the data into test and train and normalize the data to avoid bias towards any particular feature. The final step is building the model. We used four classification algorithms ( ) that best fit this dataset and predict with high accuracy. After finding the best model using accuracy and cross-validation scores, hyper parameter tuning is performed on the best model.

Model/s Development and Evaluation

Cross-validation is a statistical technique for estimating the performance of machine learning models. It is commonly used in applied machine learning to compare and select models for a given predictive modelling problem. This is because it is easier to understand and implement, and generally yields less biased capacity estimates than other methods. This procedure uses her one parameter, k. This indicates the number of groups into which a given data sample should be divided. Therefore, this procedure is often called k-fold cross-validation. If a specific value for k is chosen, it can be used in place of k in references to the model. B. k = 10 is 10-fold cross-validation.

A confusion matrix is a table commonly used to describe the performance of a classification model (or "classifier") on a set of test data whose true values are known. The confusion matrix itself is relatively easy to understand, but the associated terminology can be confusing. This is a performance measure for machine learning classification problems whose output can be more than one class. Here is a table with four different combinations of predicted and actual values. It is very useful for measuring recall, precision, specificity, accuracy and most importantly AUC-ROC curves.

Log loss is the primary classification metric based on probability. ... For a given problem, lower log loss values mean better predictions.

AUC-ROC curves are performance measures for classification problems at various threshold settings. ROC is the probability curve and AUC represents the degree or measure of reparability. Indicates how well the model can distinguish between classes. The higher the AUC, the more the model predicts the 0 class as 0 and the 1 class as 1.

Many machine learning algorithms are used for prediction. However, the predictive accuracy of these algorithms is highly dependent on the data provided during model training. If the data are in bad shape, the model will over fit and become inefficient. In short, data pre-processing is an important part of this experiment and influences the final results. Therefore, several combinations of pre-processing methods should be tested before the data is ready for training. After analysing each model, logistic regression shows good accuracy, with minimal CV difference, and hyper parameter tuning reaches 96% accuracy.

CONCLUSION

Communication is one of the basic requirements in every human life. People need to talk and interact with each other in order to express their thoughts. Over the last few years, social media and social networking have grown exponentially due to the surge in internet usage. Online conversations generate a large amount of information every day as people can discuss, speak up and express their opinions through these platforms. While this situation can be highly productive and contribute significantly to people's quality of life, it can also be destructive and extremely dangerous. Responsibility for moderating and monitoring these comments rests with your organization's social media administrator or host.

This research focused on developing a model that uses logistic regression to automatically classify comments as malicious or non-malicious. Therefore, this study aims to develop a multi-head model for detecting various types of malicious comments, such as threats, rudeness, insults, and disgust. By collecting and pre-processing malicious comments for training and testing using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm, multi-head model development uses logistic regression to generate a wide variety of to detect malicious comments, train a data set and model, and evaluate the confusion metric.

According to my understanding and research of the dataset, I presented the best model with high accuracy. This is the best model until someone comes up with an exceptional approach and technique. Furthermore, although this study does not cover all classification algorithms, we have selected Focus on algorithms.

In the future, this machine learning model can connect to various websites that can provide real-time data for price prediction. You can also add historical fare data at scale to help improve the accuracy of your machine learning model. You can create an Android app as a user interface for interacting with your users. To improve performance, we carefully design our deep learning network structure, use adaptive learning rates, and plan to train on data clusters instead of entire data sets.