Toxic Comment Classification using Classical ML Techniques and LSTM, BERT Model

Project 2 (EAS 595)

Prepared by :
Divyesh Pratap Singh
Masters in Artificial Intelligence
dsingh27@buffalo.edu

Submitted to :
Prof. Sreyasee D.B.
November, 2019
University at Buffalo, New York

November, 2023

1 Part 1: Data Preprocessing

1.1 Briefly describe the nature of your dataset?

The given dataset is a collection of comments retrieved from online sources. The comments are classified into the following categories:

Toxic	Severe Toxic	Obscene	Threat	Insult	Identity Hate

Table 1.1: Comments Categories

There are certain comments which fall into neither of the above categories, and hence are not toxic or anything from the above categories.

1.2 What is the purpose of tokenizing?

Computers cannot understand text languages in its original form. They need to be converted into mathematical vectors before PC's can make any sense or find patterns in them. Tokenization is a part of this pre-processing where sentences are split into induvial words so that later they can be converted into mathematical vectors.

1.3 What is the purpose of using TF-IDF transformer?

A simple vectorizer method to convert words corpus into mathematical vector matrices cannot understand the relationship between words. TF-IDF transformer tries to solve two problems: First it places words that are close in dictionary meaning together. Second it prioritizes those words more provides more information gain in sequences. TF-IDF suggests Term Frequency inverse Document frequency, so instead of just giving weightage as per term frequency, it gives inverse weightage to those words which are present in more documents.

1.4 Include and infer from the bar graph?

From the bar graph 'toxic' comments are the most frequently classified about 16,000 times. Although it also reflects the class imbalance as there are about 1.5 lacs words out of which only 10% are toxic. On further analysis it is also evident that all severe toxic are also marked as toxic, which indicates severe toxic is a subset of toxic. Next is obscene at about 8500 samples and then insult at 8000.

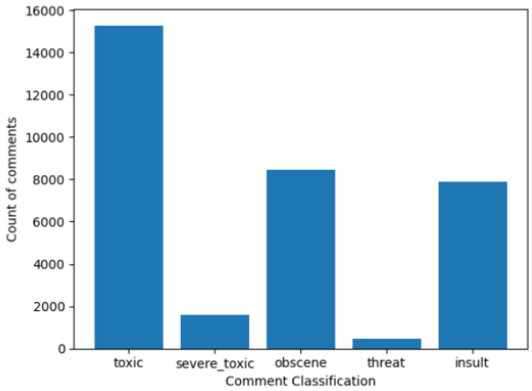


Figure 1.1: Comments Distribution

1.5 Provide the correlation matrix and write your inference about it.

From the correlation matrix, obscene and insult have high correlation with insult and vice versa. Other columns do not indicate significant correlation.

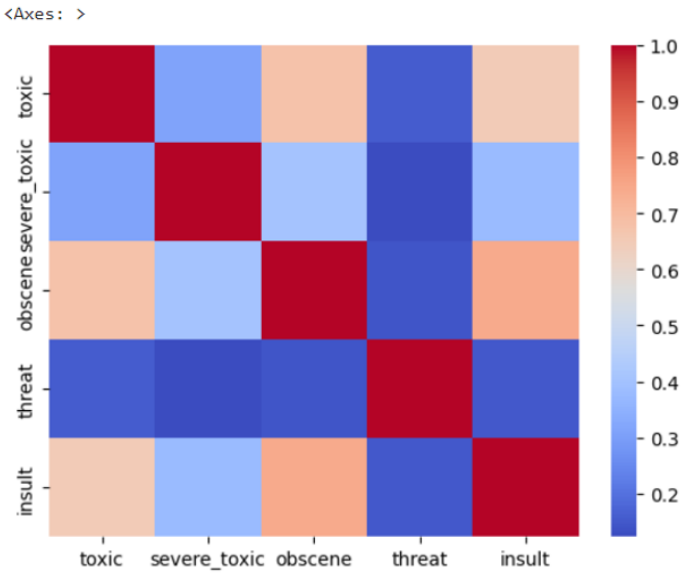


Figure 1.2: Correlation Matrix

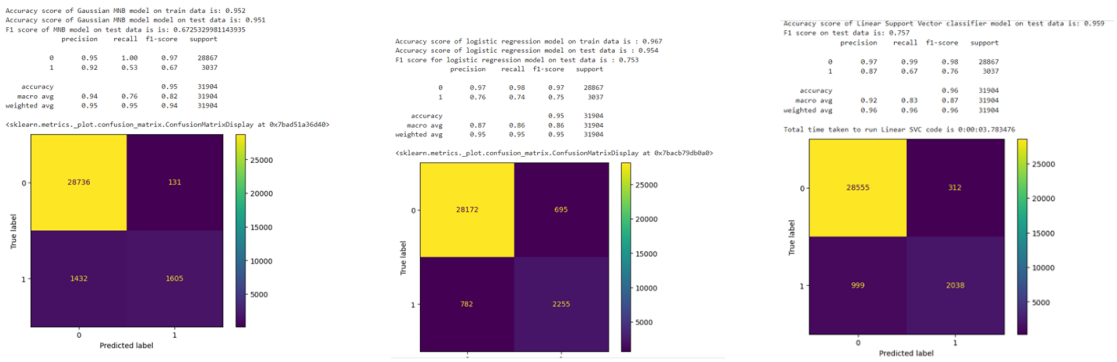
2 Part 2: Modeling and Evaluation

2.1 In two lines write your understanding of the above three classical models.

Multinomial Naïve Bayes is a version of standard Naïve bayes optimized to work on NLP tasks, while logistic regression is a standard classification model while Linear Support vector classifier is a classification model of Support vector. All three are conventional ML models not designed specifically for the NLP task.

2.2 Include all the results and plots obtained above and write your conclusion.

From the plots Linear SVC is performing best but F1 score is still not very good. On top it might look like the models are performing well since accuracy is greater than 90% for all, but it should also be noted that there is high-class imbalance (9:1). In such case accuracy is not a good parameter for model evaluation, and F1 score of low class should be checked, which is very low at about 0.6 to 0.7 for all three models.



(a) Gaussian Naive Bayes Model (b) Logistic Regression Model (c) Linear Support Vector Classifier

Figure 2.1: Multiple Models Comparison

2.3 Which model do you think performed well and why?

Out of the three models, SVM is at performing best, as the F1 score is okay, along with a good balance of precision and recall at 0.92 and 0.83 respectively.

3 Part 3: LSTM and BERT

3.1 Report all 5 custom inputs and corresponding toxicity probabilities.

S.No.	Text Input	LSTM Probability	BERT Probability
1	you bloody bastard go to hell	0.974	0.995
2	He is a good person	0.134	0.0.0217
3	You are not a bastard I am sorry	0.762	0.0.959
4	I am sorry for calling you a retard	0.635	0.0.897
5	You are a retard	0.787	0.9287
6	You are not at all a good person	0.083	0.0.049

Table 3.1: Toxic comment probability score (higher number indicates more toxic)

3.2 Do you think LSTM/ BERT or classifier models work well with these datasets?

From the custom inputs and the F1 score, LSTM model is performing better. LSTM although to a very small extent can understand sequential relation in words, as toxic class probabilities is decreasing when a negation (like 'not') is being used., as seen from example 1 and 3 difference.

References

1. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
2. https://www.youtube.com/watch?v=ZUqB-luawZg
3. https://medium.com/analytics-vidhya/getting-started-with-nlp-tokenization-document-term-matrix-tf-idf-2ea7d01f1942
4. https://www.nltk.org/api/nltk.tag.pos_tag.html
5. https://www.nltk.org/_modules/nltk/stem/wordnet.html
6. https://www.geeksforgeeks.org/python-lemmatization-with-nltk/
7. https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM
8. https://datascience.stackexchange.com/questions/72296/predict-proba-for-binary-classifier-in-tensorflow

Report END