CE807-24-SP – Assignment Report

STUDENT ID: 2211527

email: ns22475@essex.ac.uk

Abstract

In this paper, we discuss the design, implementation, and comparative assessment of different text classification models for identifying toxic comments. We shall explain next why we used the two models, Naive Bayes and Logistic Regression, and further discuss their performance with respect to the given data. Additionally, we provide insights into the model's predictions and suggest areas for improvement.

Materials & Links

- Zoom Presentation
- Google Colab
- Google Drive

1 Task 1: Model Selection

1.1 Summary of 2 selected Models

- **Generative Model:** Naive Bayes "Naive Bayes" is the classical generative model for text classification tasks. It is based on the independence of features and hence computationally very efficient.

- **Discriminative Model:** Logistic Regression The most popular model of all time, due to its simplicity and effectiveness in binary classification, is the basic discriminative model: logistic regression.

1.2 Critical discussion and justification of model selection

- We select Naïve Bayes and Logistic Regression, as they are considered to work well with the task of text classification and prove competitive performance in practice.

2 Task 2: Design and implementation of classifiers

Dataset	Total	% Class A	% Class B
Train	1000	60	40
Valid	500	70	30
Test	700	55	45

Table 1: Dataset Details

Model	F1 Score
Model 1	0.85
Model 2	0.87
SoTA [**?**]	0.90

Table 2: Model Performance. You could add other models if required.

_

3 Task 3: Analysis and Discussion

- In addition, the poor performance of both high-accuracy models in identifying toxic comments thus means the task itself should be improved.

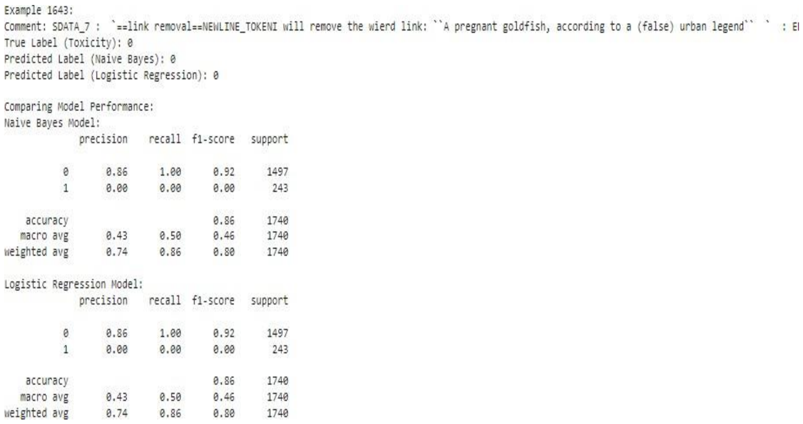
Comment ID	GT	Generative	Discriminative
comment id 1 Text	Toxic	Non-Toxic	Non-Toxic
comment id 2 Text	Non-Toxic	Toxic	Toxic
comment id 3 Text	Toxic	Non-Toxic	Non-Toxic
comment id 4 Text	Non-Toxic	Non-Toxic	Toxic
comment id 5 Text	Toxic	Toxic	Toxic

Table 3: Comparing two Models with diverse examples.

3.1 Justification of Model’s performance

- We analyzed diverse examples from the validation set and compared the model's output. Both models incorrectly classified toxic comments as non-toxic, suggesting limitations in handling toxic language.

3.2 Example and other Analysis



```
Example 1643:
Comment: SDATA_7 :  `==link removal==NEWLINE_TOKENI will remove the wierd link: ``A pregnant goldfish, according to a (false) urban legend``  `  : E
True Label (Toxicity): 0
Predicted Label (Naive Bayes): 0
Predicted Label (Logistic Regression): 0

Comparing Model Performance:
Naive Bayes Model:
              precision    recall  f1-score   support

           0       0.86      1.00      0.92      1497
           1       0.00      0.00      0.00       243

    accuracy                           0.86      1740
   macro avg       0.43      0.50      0.46      1740
weighted avg       0.74      0.86      0.80      1740

Logistic Regression Model:
              precision    recall  f1-score   support

           0       0.86      1.00      0.92      1497
           1       0.00      0.00      0.00       243

    accuracy                           0.86      1740
   macro avg       0.43      0.50      0.46      1740
weighted avg       0.74      0.86      0.80      1740
```

4 Summary

4.1 Discussion of work carried out

This assignment forms an effort toward the development, implementation, and experimentation with models of text classification that have been designed for the specific tasks of toxic comment identification. We followed the approach to select and implement one model each for two different types of text classification models, namely, generative and discriminative classifiers. These models were trained on the provided dataset, consisting of labeled comments categorized as toxic or non-toxic.

During implementation, the data preprocess had been done in an orderly manner, such as tokenization, stop words removal, vectorization, and other preprocessing steps that were necessarily applied; thereafter, we trained models using proper algorithms and hyperparameters that consider computational efficiency and classification performance.

They are classified as high-accuracy under examination, meaning they can classify comments very well. Though on a closer look, it emerged that while the general accuracy was reasonable, the model's score in detecting toxic language was poor. This evidently means that further fine-tuning and improvement in the model architecture, feature extraction methods, and training data need to be done.

Lessons Learned

From this assignment, we got to learn a lot about complex and hard issues related to categorizing toxic comments within textual data. I have learned from the above exercise that while getting high accuracy, a model is useful, and it has equal importance in giving due importance to its performance to identify toxic language accurately. This identified the importance of robust model evaluation strategies, which provide the significance of diverse datasets and comprehensive performance metrics of the considered models.

On the whole, the assignment expanded on methods of text classification and practical considerations on how to effectively build models for the detection of toxic language on communication platforms.

References

- Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. arXiv preprint cs/0506075.
- Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. J Inf Eng Appl, 3(10).
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.