REPORT

CSE712: SYMBOLIC MACHINE LEARNING II

SUBMITTED TO:

Department of Computer Science and Engineering
BRAC University

SUBMITTED BY:

SAKIB ROKONI

Master of Engineering in
Computer Science and Engineering
BRAC University.

Explainable Detection of Online Sexism

Introduction:

Online platforms have enabled significant worldwide communication and involvement in recent years. Online sexism, where biased behaviors and opinions appear in many digital formats, has also expanded with connectivity. Outright harassment and subtle macroaggressions hamper inclusive and egalitarian online spaces. Combating online sexism requires knowing its causes and spotting it. Effective treatments and good behavior changes require understanding and interpreting detection data. Investigating digital misogyny involves exposing and explaining.

Identifying and analyzing sexist conversations online is explained by its methodology, obstacles, and results. This study discusses automated detection systems' detection techniques, machine learning, and natural language processing and their ethical consequences. Understanding the challenges of detecting and describing digital sexism will help us develop ethical and effective solutions. We want to make the internet fair and more inclusive.

Motivation:

The ubiquitous nature of digital platforms makes it all the more important to address online misogyny, which violates basic ideals of equality and respect. It has an impact on offline social attitudes and harms people's safety and well-being, especially members of marginalized groups. Ensuring safe and inclusive digital environments is of the utmost importance, particularly for vulnerable groups such as children and adolescents, due to the exponential rise of online platforms. Machine learning and natural language processing are two examples of how technological progress has opened up exciting new possibilities for the detection and elimination of widespread internet misogyny. Creating an environment where everyone feels welcome and valued is the ultimate goal of these initiatives.

Dataset:

For this assignment the dataset named, train_all_tasks.csv has been used. This dataset consists of text examples with labels related to sexism. A rewire_id is used to uniquely identify each instance, probably for organizational reasons. The material itself is found in the 'text' column, whereas 'label_sexist' designates whether or not the text is sexist.A 'label_category' column also suggests classification that goes beyond sexism. Finally, vectored versions of labels may be represented by the 'label_vector' column, perhaps for machine learning uses. More information about the quantity, context, and possible uses of the dataset would give a more thorough picture.

Loading and Exploring Data:

Several libraries were used for text vectorization as part of the text preprocessing before training the model, data cleaning, and EDA.

It is verified for the data cleaning stage whether any nan or null values should be dropped. After that, I worked on the text preprocessing section, making all of the words lowercase and checking the Regular Expressions, Capitalization, double spaces, and stop words to be eliminated from the dataset. Here the NLTK library has been utilized to carry out the tasks.
For instance, I looked up statistical analysis in the EDA section and examined the class distributions as well as the most often used words in two categories—sexist and non-sexist words. Consequently, I examined the texts' average length.
Apart from that, many visualizations have been carried out with different Python libraries including matplotlib, seaborn numpy, and pandas.

MODELs Selection and Implementation:
To complete Task A, I implemented several classification models, including AdaBoost, Random Forest, and SVM. Prior to that, the text vectorization method known as TFIDF vectorization was utilised in NLP. In the case of Task B, a model based on neural networks performed admirably with improved precision. The accuracies of the separate models are as follows: 82% for Random Forest, 82.89% for SVM, 71% for LSTM, and 82% for AdaBoost.
Using classification reports and confusion matrices, each model result has been thus represented.
The justification for employing each of these models is that they are all well-known and widely used for text classifications, including both traditional and neural network-based models that exhibit superior performance as an accuracy metric.

Results:
For Task A
Binary Sexism Detection: a two-class (or binary) classification where systems have to predict whether a post is sexist or not sexist.

Accuracy of Random Forest Classifier: 0.8228571428571428
Classification Report:

	precision	recall	f1-score	support
not sexist	0.82	0.98	0.89	2096
sexist	0.88	0.34	0.49	704
accuracy			0.82	2800
macro avg	0.85	0.66	0.69	2800
weighted avg	0.83	0.82	0.79	2800

Confusion Matrix:
[[2064 32]
[464 240]]

Accuracy of SVM Classifier: 0.8289285714285715
Classification Report:

```
              precision    recall  f1-score   support

  not sexist       0.83      0.98      0.90      2096
      sexist       0.85      0.39      0.53       704

    accuracy                           0.83      2800
   macro avg       0.84      0.68      0.71      2800
weighted avg       0.83      0.83      0.80      2800


Confusion Matrix:
[[2046   50]
[ 429  275]]
```

Accuracy of AdaBoost Classifier: 0.825

```
Classification Report:
              precision    recall  f1-score   support

  not sexist       0.83      0.96      0.89      2096
      sexist       0.77      0.44      0.56       704

    accuracy                           0.82      2800
   macro avg       0.80      0.70      0.72      2800
weighted avg       0.82      0.82      0.81      2800


Confusion Matrix:
[[2002   94]
[ 396  308]]
```

From the comparison, it seems SVM has slightly higher accuracy and better precision-recall balance compared to the other classifiers.

For Task B

Category of Sexism: for posts which are sexist, a four-class classification where systems have to predict one of four categories: (1) threats, (2) derogation, (3) animosity, (4) prejudiced discussions.

LSTM Accuracy: 0.7114285714285714

```
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.88      0.86      2096
           1       0.04      0.02      0.03        41
           2       0.17      0.14      0.16       140
           3       0.03      0.04      0.03        48
           4       0.48      0.54      0.51       142
           5       0.23      0.09      0.13        53
```

6	0.22	0.19	0.21	121
7	0.11	0.11	0.11	96
8	0.00	0.00	0.00	13
9	0.00	0.00	0.00	18
10	0.00	0.00	0.00	20
11	0.00	0.00	0.00	12
accuracy			0.71	2800
macro avg	0.18	0.17	0.17	2800
weighted avg	0.68	0.71	0.70	2800

Limitation:

Online sexism detection has difficulties in effectively identifying nuanced and context-dependent sexist language and behavior, as context is critical. Furthermore, biases in training data might result in unjust predictions, particularly for minority or marginalized groups, repeating past injustice. The dynamic nature of internet communication is a difficulty, necessitating constant updating to detect emerging kinds of sexism. Furthermore, balancing privacy and safety is crucial, ensuring that detection methods respect individuals' rights to free expression and privacy.

Future work:

To successfully detect nuanced sexist discourse online, creating new algorithms combining natural language processing and machine learning is needed. To avoid unfair predictions, training data biases and algorithmic fairness must be addressed. Continuously monitoring and adapting detection systems to online communication landscapes is important. For safer and more inclusive digital spaces, research into effective intervention tactics and regulations that respect individual rights and privacy is crucial.