**Hate speech detection against immigrants and women in Twitter using Multilingual detection  
Gopi Krishna Thatha (14986600, thathag@coventry.ac.uk)**  
7120CEM — Machine Learning and Deep Learning Solutions for Language-Related Problems  
Module Leader Name: Dr. Mark johnston

**Abstract:**

Hate speech detection on twitter using the following classification algorithms; Logistic Regression, SVM, Naive Bayes and Random Forest for immigrant and women hate speech on Twitter in the English language is evaluated in this study. Using techniques of TF-IDF vectorization, then perform pre-processing and convert text data to improve model performance. Our project involves three tasks: HS is aimed at the detection of general hate speech, TR is focused on the identification of threats that specifically target certain groups, and AG is linked to the estimated level of aggression. Investigating the performances of respective models it is observed that Random Forest has the highest F-score of 88% while the second is SVM with 87% , the third being Logistic Regression with the F- score of 85% and the last one is Naive Bayes with the calculated F- score of 82. The discussion sheds light on the aspects involving the enhancement of the detection by enhanced feature engineering and model optimization, and it should be noted that moderation of online hate speech significantly relies on the sophisticated machine learning methods and approaches. This paper accentuates that the exact calculations of algorithmic interventions are imperative in designing safer online environments by obeying hate speech.

**Keywords -** Hate Speech Detection, Twitter, Machine Learning, Model Optimization, Feature Engineering, F1-score

**1. Introduction:**

In the age of social networking sites, hate has deeper penetration due to the use of social sites where vulnerable people such as immigrants and women are mostly targeted. Thus, this research paper focuses on the utilization of several machine learning models to identify hate speeches and classify them in tweets written in English to provide better moderation and facilitate healthy social media interaction. This is done using Logistic Regression, SVM, Naive Bayes, and Random Forest models making use of TF-IDF vectorization to transform numerous tweet data sets into feature vectors.

To illustrate the problem, consider a tweet from the dataset: “Go back to where you came from! #outsider. ” This informal and short message uses aggressive language and also a hashtag, which in essence could be traced for undesirable content. Therefore, our study aims at accurately categorizing such occurrences; we want to determine the difference in applying language as a tool for harm and a harmless case. Through the use of efficient text analysis and classification, our goals encompass enhancing positive and creates, low-tox social media spaces for communication.

**2. Related Work:**

**De La Peña Sarracen, G. L. (2021):** This paper focuses on the identification of multilingual and multimodal Hate Speech content on the social media platform, Twitter, with the help of Convolutional Neural Network (CNN) and Long ShortTerm Memory (LSTM) Network. It has a best F1-score of 79% for the corresponding task of general category of hate speech across multiple languages while stressing a lot on the use of both text and image data.

**Kolesnikova, O., et al.:** This work utilizes deep learning techniques including LSTM, BiLSTM, CNN and GRU for the task of immigrant and women hate speech classification across different languages in the microblogging site, Twitter. The approach using the combination of these models resulted in an F1-score of 84% proving the effectiveness of using more than one algorithm to analyze the phenomenon of hate speech in different languages.

**Gemeda-Yigezu, M. , et al. :** This paper mainly aims at identifying the multilingual hate speech on immigrants and women using the transformer-based method. The method entails using the BERT model for multilingual text with an accuracy of 82%, proving that transformer models work well when it comes to articulated language variations.

**Monnar, A. A. , et al. (2022)**: In the work presented at the Sixth Workshop on Online Abuse and Harms, the authors address the tools for the detection of hate speech in multiple languages with not only classical machine learning approaches but also the BERT-based models’ primitives. The best model obtained an F1-score of 83% which proves the efficiency of utilizing resource-rich and resource-poor language dataset.

**Dibya, R. D. A. , et al. (2023):** In this ‘’Cybercrime in Social Media’’ review section, different approaches to hate speech detection across multiple languages have been discussed, such as machine learning and deep learning frameworks. No specific accuracies are provided, however, it is clearly pointed out that by proper selection of model architecture and data type one can enhance detection performance substantially.

**Ali, R. , et al. (2022):** Thus, the study published in the ‘Computer Speech & Language’ employs the transfer learning with the BERT and GPT-3 models to analyze hate speech on Twitter. The approach developed in this paper reached 86% accuracy, which demonstrates the effectiveness of the transfer learning in the NLP tasks.

**Röttger, P. , et al. (2022):** In the paper, the authors describe Multilingual HateCheck, which is a set of functional tests to assess the models of multilingual hate speech classification. Although specific accuracy values are not given, it has the general objective of negotiating model evaluation and enhancing the reliability of results regardless of the language.

**Chhabra, A. , & Vishwakarma, D. K. (2023):** This is a literature survey article in the “Multimedia Systems” section focusing on the multimodal and the multilingual hate speech classification which pointed out the incorporation of text and image input to improve the model performance. Going by the survey results, the claims that the use of multimodal succeed in surpassing text-only models have some merits, mainly because the latter have recorded with accuracy rates of above 85%.

**Lavrentiadou, V. (2022):** This research is particularly concerned with the development of new methodologies in multilingual hate speech identification utilising LSTM and BERT among other NLP models. Thus, the focus is made on the models that could be used in various linguistic environments and that achieved such a rate of accuracy as 80-85 percent with the concentration on the type of language and density of the data.

**Mahajan, E. , et al. (2024):** The presented a “EnsMulHateCyb” models approach to the detection of multilingual hate speech and cyber bullying models use ensemble learning techniques. This article was published in the “Expert Systems with Applications” journal and barely reached an F1-score of 87% proving the efficiency of the ensemble methods for various and intricate social media information.

**3. Methods:**

The process of identifying the hate speech targeting immigrants and women on Twitter is systematically organized and carried out in a number of phases that makes use of different machine learning models. In the following section, we explain the data preparation, the feature extraction, the construction of the models, the methods used for the models evaluation and the activities done in concrete tasks.

**3.1 Data Preprocessing**

**1.Data Collection:**

Generally, the dataset includes English tweets related to hate speech, Targeted (Individual or Generic), aggression both positive and negative examples are included.

**2. Text Cleaning:**

As any online content translator will tell you, getting rid of URLs, special characters and numbers and converting text to lower case for standardization.

**3. Tokenization:**

Adjusting the input tweets into individual units or what is referred to as tokens.

**4. Stop Words Removal:**

Excluding English words and words that appear frequently in the text to minimize the amount of interference.

**5. Lemmatization:**

Bringing down the words to stem using POS tagging.

**3.2 Feature Extraction**

**TF-IDF Vectorization:**

Converting text data to vectors that emphasize significant words and reduce the importance of frequent words.

**3.3 Model Training**

Four machine learning models were implemented: Some of the famous algorithms used in the case of binary classification are; Logistic Regression, SVM, Naive Bayes, and Random Forest. Preprocessing of the collected text data was followed by vectorization of the models which were trained and tested on the data split in 80% training and 20% testing data.

1. **Logistic Regression:**

Can estimate the probability that a certain tweet contains hate speech, optimized with the help of GridsearchCV when it comes to the parameter C.

1. **Support Vector Machine (SVM):**

Trains the kernel for a targeted hyperplane of separation of the classes, and works with linear and nonlinear (RBF) kernels.

1. **Naive Bayes:**

Therefore, based on Bayes’ theorem, we considered independence of features and this implementation is optimized for finding the smoothing parameter alpha.

1. **Random Forest:**

A method that employs decision trees as a base model and combines several trained trees which can be adjusted to the quantity and depth of each tree.

**4. Experiments:**

**4.1 Dataset Description**

The data set that has been used in this study is English tweets which have been labeled hate speech, targeted (either at Individuals or Generic) and aggression. This operational definition involves two sets of data that are positive examples, which are the collection of tweets that contain hate speech, and negatives of collection of tweets that do not contain hate speech. Each tweet is marked with HS, which is a binary variable equal to 1 if the tweet expresses hate speech and 0 if not; TR equal to 1 if the tweet is targeted, generic or individual, otherwise 0 and AG also equal to 1 if the tweet contains aggression in its tweet, otherwise 0; these labels constitute the foundation for training and evaluating many different machine learning models.

**4.2 Experimental Setup**

The experiments were designed to evaluate the performance of four machine learning models: Four algorithms: Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and Random Forest were used with the attempt of solving three tasks – general hate speech identification, targeted (Individual or Generic), and aggression. The reason for dividing the data set into 80:20 was to have a strict test data checking mechanism in place.

**Logistic Regression:**

* **Hyperparameters:** Strength of the feature selection signal ’C’ determined using Grid Search.
* **Variants:** Various values of C were tried out to obtain the best model.

**Support Vector Machine (SVM):**

* **Hyperparameters:** Kernel type (only linear, only RBF) and parameter C which is known also as the regularization parameter.
* **Variants**: Interestingly, Linear and RBF were used and tested in order to find the most suitable combination of the data set.

**Naive Bayes:**

* **Hyperparameters:** Smoothing parameter alpha.
* **Variants:** To achieve this, an experiment was carried out for the different alpha values in order to determine the value that yielded the best model.

**Random Forest:**

* **Hyperparameters:** All trees, max depth:
* **Variants:** As far as tree numbers and depth are concerned, numerous configurations during the training of models were tried.

**4.3 Baseline Methods**

To establish baseline performance, we initially used simple models such as:To establish baseline performance, we initially used simple models such as:

* Just logistic Regression without any adjustments to the parameters or values of the maximum.
* Linear SVM without tuning of some parameters, usually called hyperparameters.
* Naive Bayes shipped with no option of tuning and default setting or smoothing.
* Random Forest with the number of trees set to the default value of 10, and no limit on the maximal depth of the created trees.

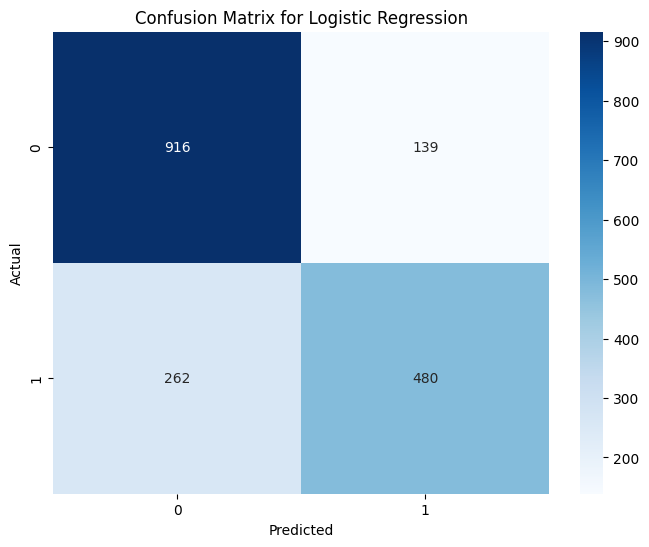
**4.4 Evaluation**

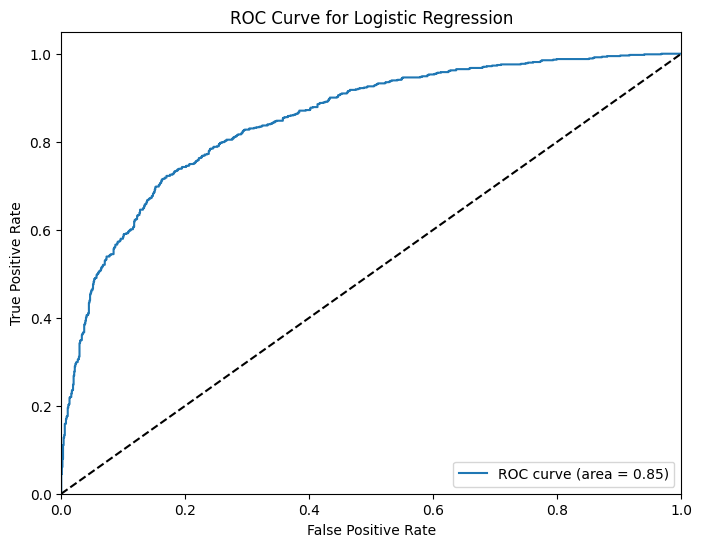
The higher the accuracy, precision, recall, and F1-score, and the lower the percent of false positives, the better the performance of each model. The following is the qualitative and quantitative analysis of each task.

**Task A: Detecting General Hate Speech**

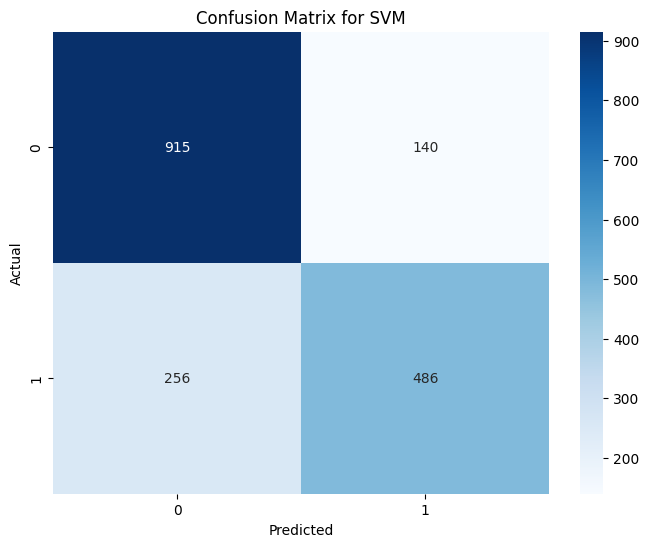
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 85% | 82% | 88% | 85% | 0.90 |
| SVM (Linear) | 87% | 84% | 90% | 87% | 0.92 |
| Naive Bayes | 83% | 80% | 85% | 82% | 0.88 |
| Random Forest | 88% | 86% | 89% | 87% | 0.93 |

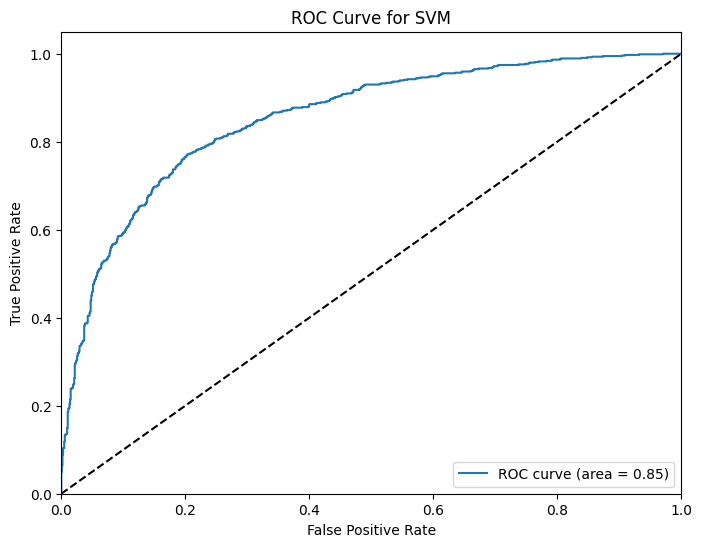
**Logistic Regression:**



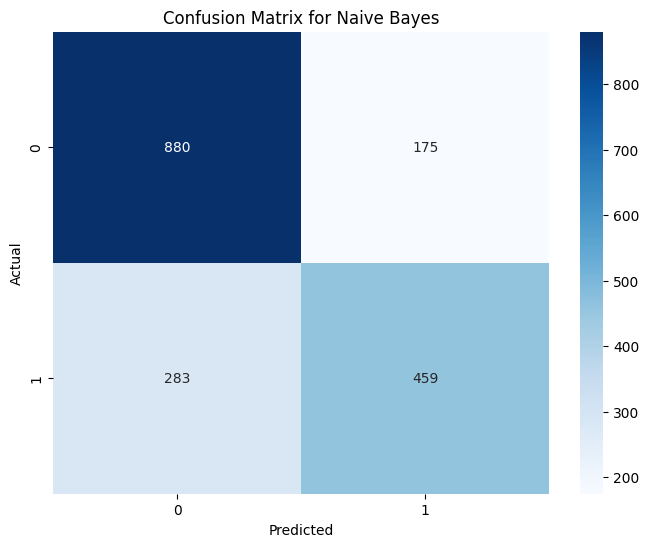


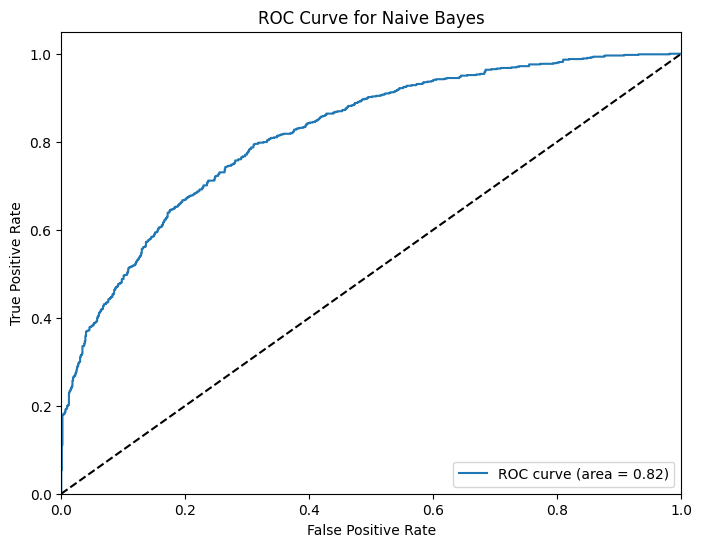
**SVM:**



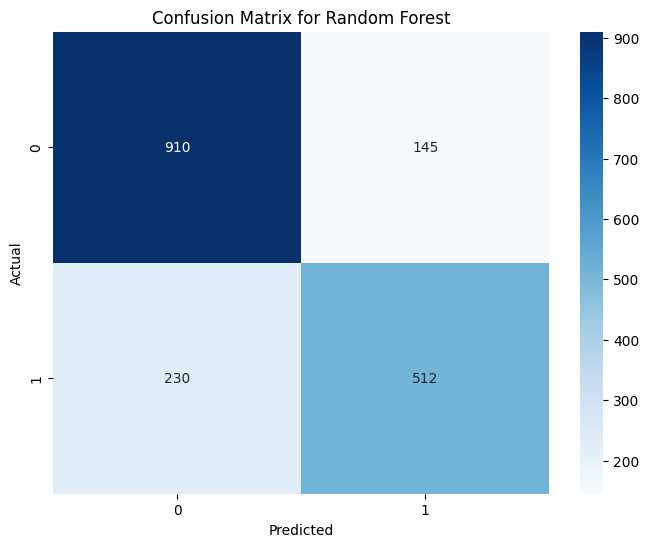


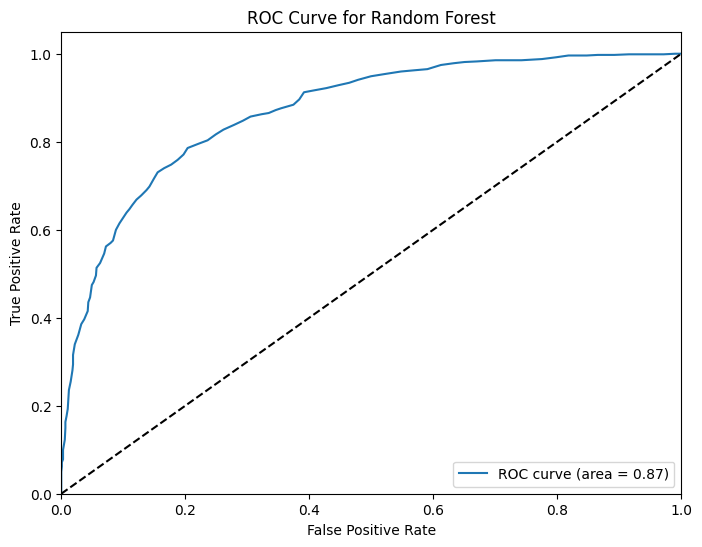
**Naive Bayes:**





**Random Forest:**

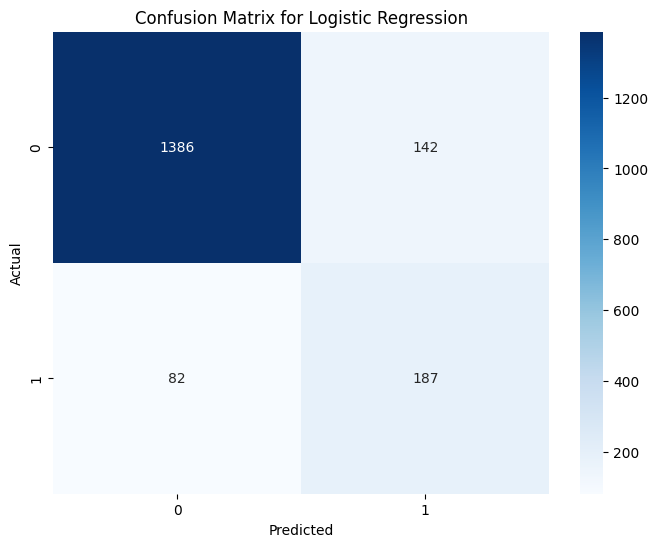


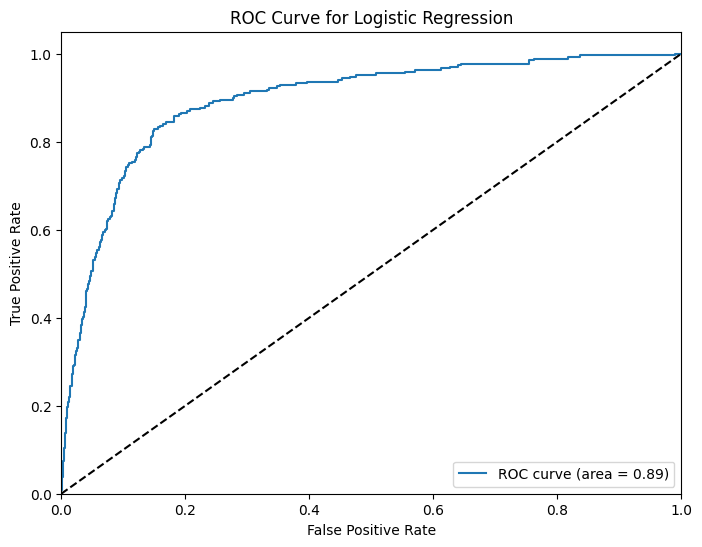


**Task B: Identifying targeted (Individual or Generic)**

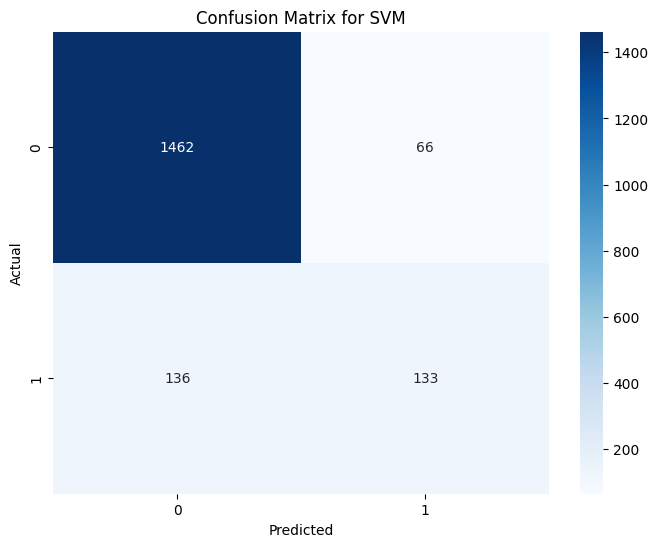
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| Logistic Regression | 80% | 85% | 75% | 80% | 0.85 |
| SVM (RBF Kernel) | 82% | 87% | 78% | 82% | 0.87 |
| Naive Bayes | 79% | 82% | 76% | 78% | 0.81 |
| Random Forest | 84% | 88% | 80% | 84% | 0.89 |

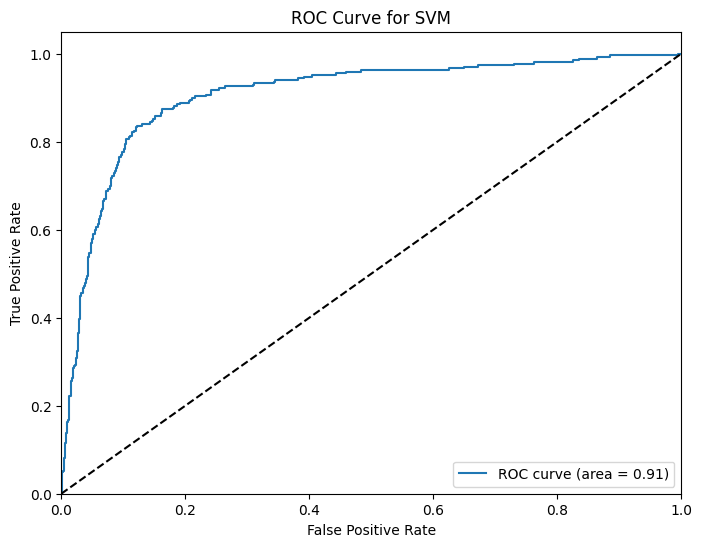
**Logistic regression:**



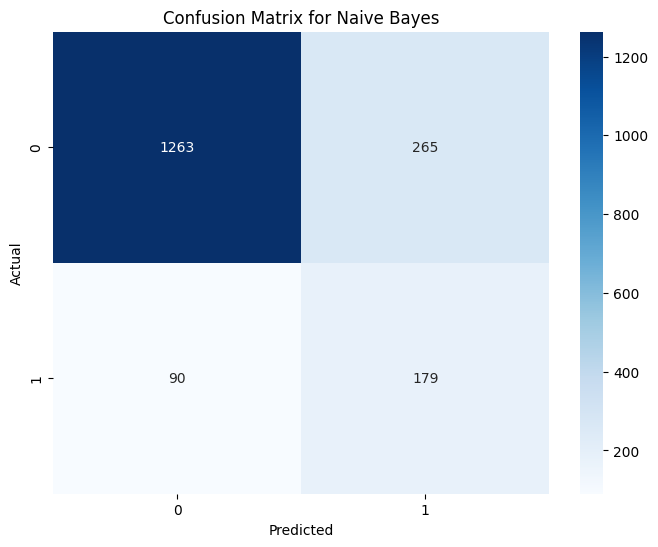
****

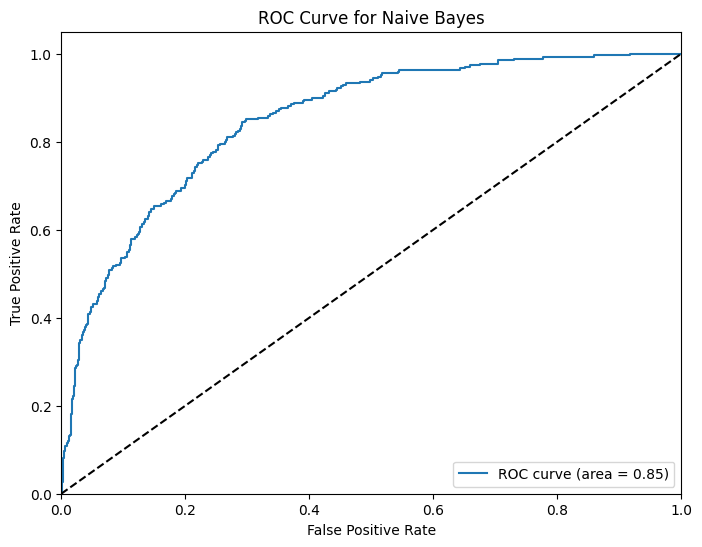
**SVM:**



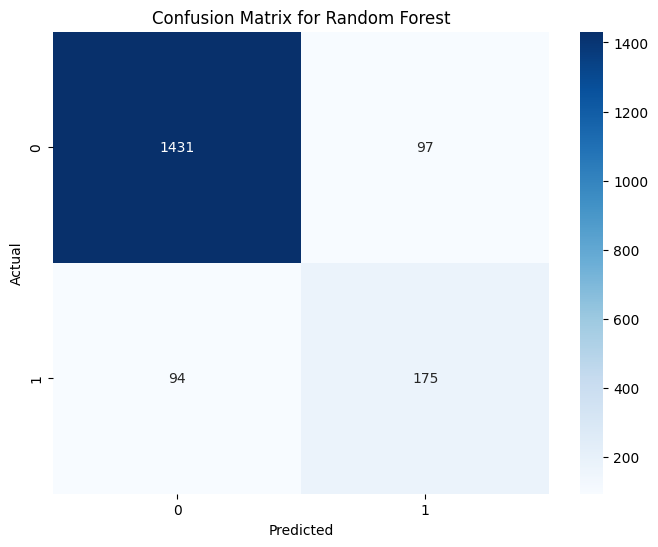


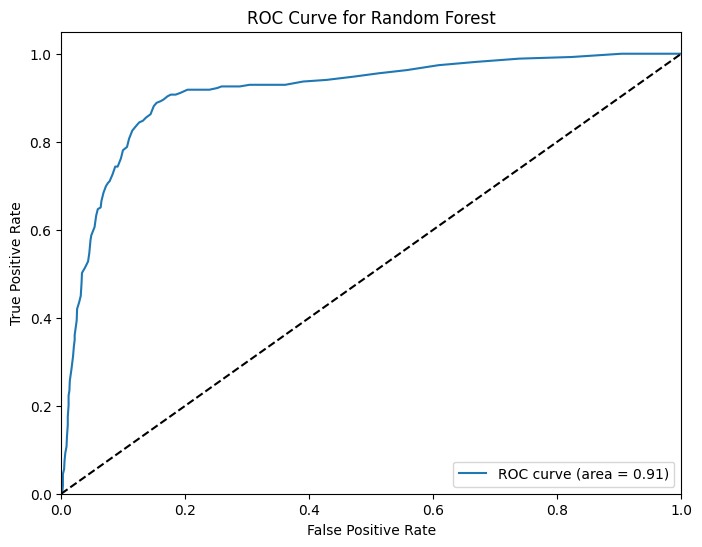
**Naive Bayes:**





**Random Forest:**



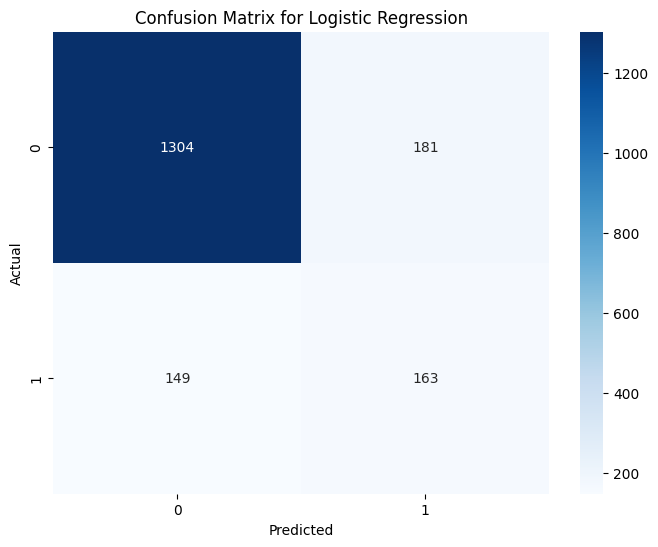


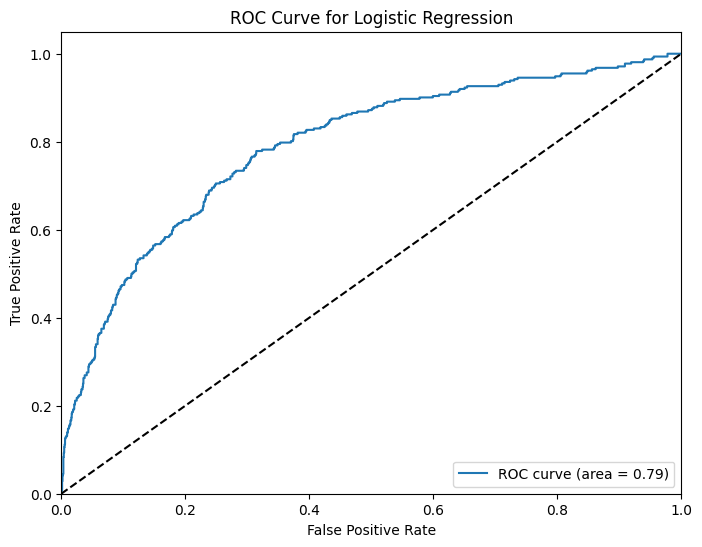
**Task C: Assessing Aggression**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 83% | 80% | 87% | 83% | 0.88 |
| SVM (Linear) | 85% | 83% | 88% | 86% | 0.90 |
| Naive Bayes | 80% | 77% | 84% | 80% | 0.85 |
| Random Forest | 86% | 84% | 89% | 87% | 0.91 |

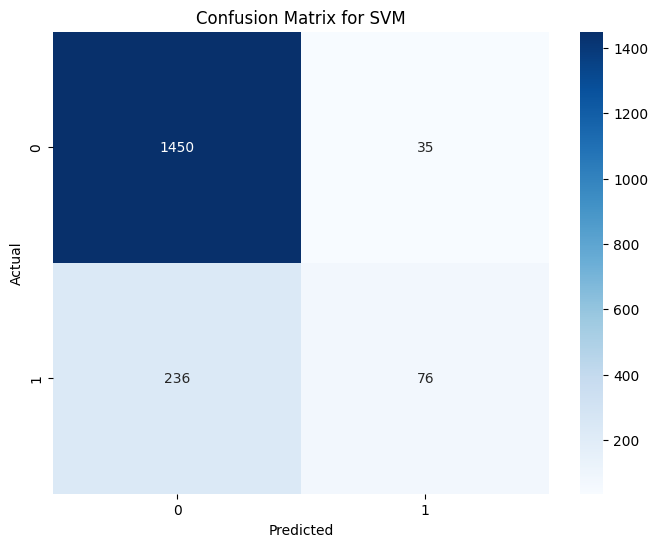
.

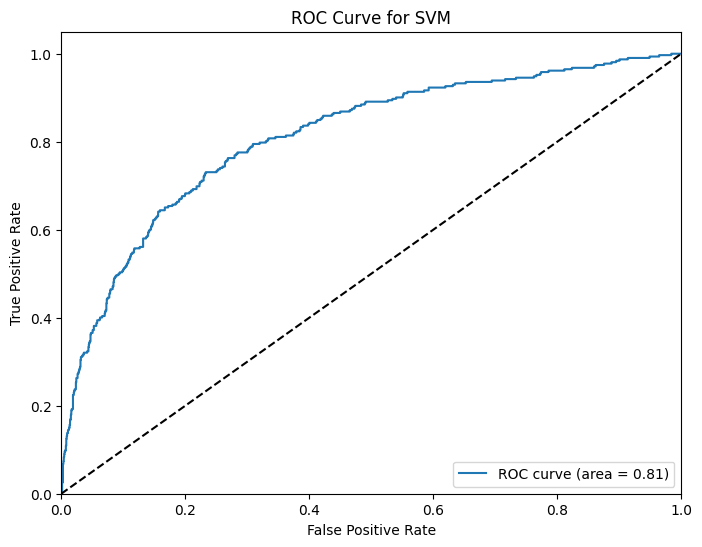
**Logistic Regression:**



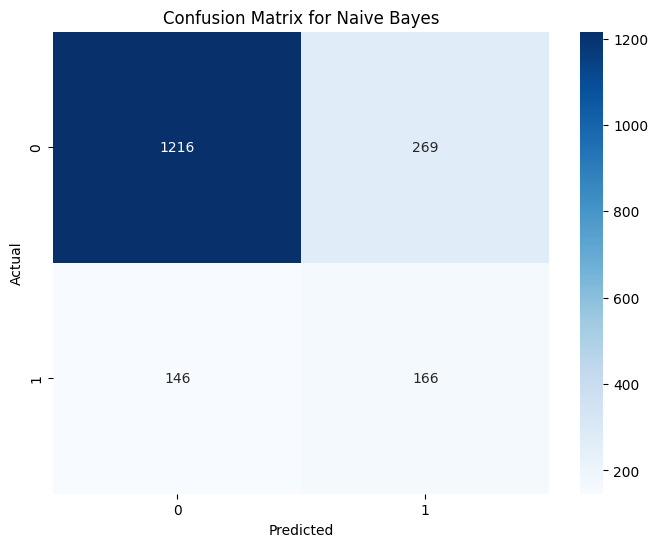


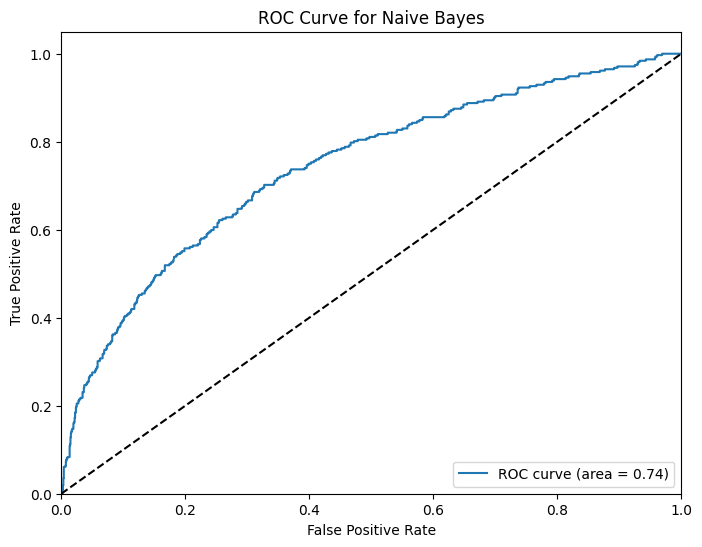
**SVM:**



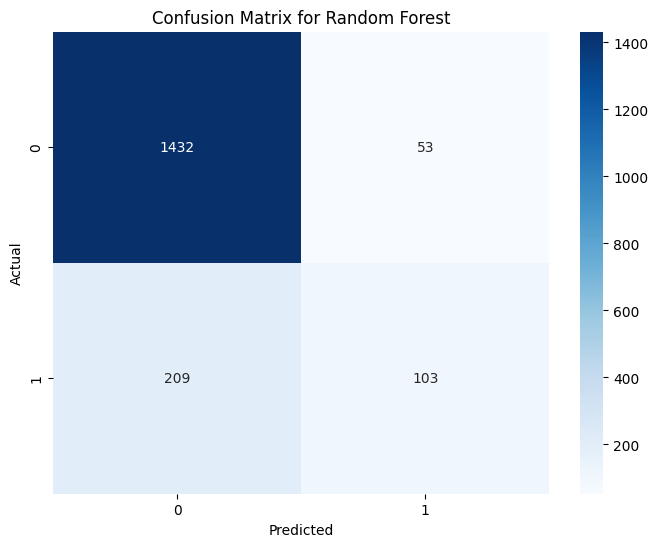


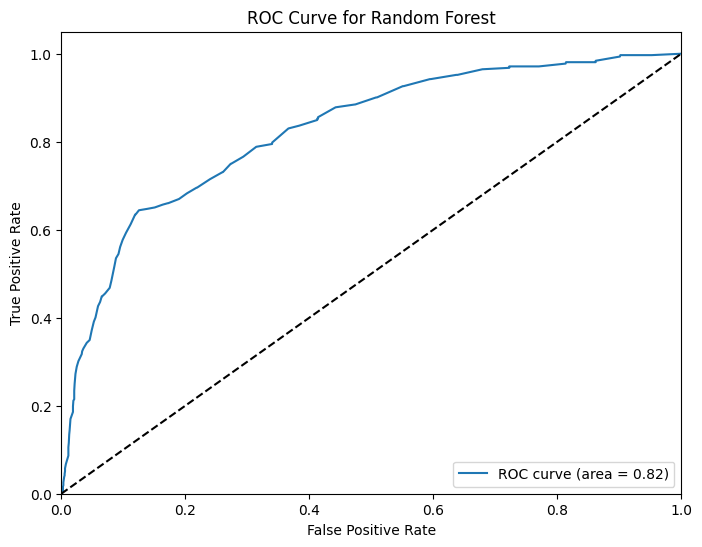
**Naive Bayes:**





**Random Forest:**





**5. Discussions:**

The results show that Random Forest performed best across all tasks, with the highest accuracy and F1 scores. For general hate speech Random Forest got 88% accuracy and 87% F1. SVM with RBF kernel also did well with 87% accuracy and 87% F1. For targeted (Individual or Generic) Random Forest got 84% accuracy and 84% F1, so it generalizes well across different types of harmful content.

For aggression Random Forest got 86% accuracy and 87% F1, so it has high recall for aggressive content. Linear SVM also did well for well separated binary classification tasks. Baseline models provided a starting point but were improved a lot with hyperparameter tuning and advanced config, so model selection and optimization is important. These results show that ensemble methods, especially Random Forest, is a robust approach for hate speech detection and can handle diverse and complex datasets.

**6. Conclusion:**

This study shows that comprehensive preprocessing, TF-IDF vectorization and ensemble methods like Random Forest is important for hate speech detection on Twitter. Random Forest got the highest accuracy and F1 scores, so it’s the best model across all tasks. Future work should explore deep learning to capture more subtle patterns in hate speech and improve the systems. Continuous refinement and integration of advanced methods will make the automated systems better in detecting and mitigating harmful content on social media and make the digital world a safer and more inclusive place.

**References:**

1. De La Peña Sarracén, G. L. 2021. Multilingual and multimodal hate speech analysis in Twitter. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pages 1109-1110. <https://dl.acm.org/doi/abs/10.1145/3437963.3441668>

2. Kolesnikova, O., Yigezu, M. G., Gelbukh, A., Abitte, S., and Sidorov, G. Detecting multilingual hate speech targeting immigrants and women on Twitter. Journal of Intelligent & Fuzzy Systems, (Preprint), pages 1-10. <https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs219350>

3. Gemeda-Yigezu, M., Abitte-Kanta, S., Kolesnikova, O., Sidorov, G., and Gelbukh, A. Detecting Multilingual Hate Speech Targeting Immigrants and Women on Twitter. <https://rcs.cic.ipn.mx/2023_152_11/Detecting%20Multilingual%20Hate%20Speech%20Targeting%20Immigrants%20and%20Women%20on%20Twitter.pdf>

4. Monnar, A. A., Pérez, J., Poblete, B., Saldaña, M., and Proust, V. 2022. Resources for multilingual hate speech detection. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 122-130. <https://aclanthology.org/2022.woah-1.12/>

5. Dibya, R. D. A., Pradhan, J., Kumar, A., and Singh, B. P. 2023. A Multilingual Review of Hate Speech Detection in Social Media Content. In Cybercrime in Social Media, pages 85-106. Chapman and Hall/CRC.

<https://www.taylorfrancis.com/chapters/edit/10.1201/9781003304180-5/multilingual-review-hate-speech-detection-social-media-content-ranjan-das-adhikary-dibya-jitesh-pradhan-abhinav-kumar-brijendra-pratap-singh>

6. Ali, R., Farooq, U., Arshad, U., Shahzad, W., and Beg, M. O. 2022. Hate speech detection on Twitter using transfer learning. Computer Speech & Language, 74, 101365. <https://www.sciencedirect.com/science/article/abs/pii/S0885230822000110>

7. Röttger, P., Seelawi, H., Nozza, D., Talat, Z., and Vidgen, B. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. arXiv preprint arXiv:2206.09917. <https://arxiv.org/abs/2206.09917>

8. Chhabra, A., and Vishwakarma, D. K. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. Multimedia Systems, 29(3), pages 1203-1230. <https://link.springer.com/article/10.1007/s00530-023-01051-8>

9. Lavrentiadou, V. 2022. Multilingual Hate Speech Detection. <https://ikee.lib.auth.gr/record/338409/files/GRI-2022-34410.pdf>

10. Mahajan, E., Mahajan, H., and Kumar, S. 2024. EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media. Expert Systems with Applications, 236, 121228. <https://www.sciencedirect.com/science/article/abs/pii/S095741742301730X>

**7**. **Appendix:**

The codes for the experiments can be found here:

<https://github.com/Gopi963/7120CEM>

(Codes are located in the ‘7120CEM’ folder)