**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:**

This research evaluates deep learning models including LSTM in identifying hatred of immigrants and women on Twitter and this is undertaken on material in English. Through the use of the embedding layer and LSTM, textual data is preprocessed and transformed for higher classification rates. Our project involves three tasks: identify general hate speech (HS), targeted (Individual or Generic) (TR), and aggression (AG). Performance analysis reveals varying effectiveness: , In this study the LSTM model attained accuracies of 70. 83% for HS, 82. For AG the control totals 83% while overall control was at 86. 67% for TR. It also identifies the possibilities of increasing detectability by enhancing feature extraction and optimizing the models. Thus, we stress on the role of advanced deep learning methods to counter online hate speech and encourage subsequent studies for their improvement to attain better online communication environments. The present paper emphasises algorithmic modulations insofar as they are conducive to controlling the flow of contents on social media, thus striving to establish the principles of safe online communities and prevent the appearance of hate speech.

**Keywords -** Hate Speech Detection, LSTM, Twitter, Deep Learning, Classification Accuracy

**1.Introduction:**

It is sad that with the increased expansion of social media sites, hate speech is being echoed more than before and in the process immigrants, women are among the most degraded. This paper is a study of the use of deep learning models in the identification and classification of hate speech in the chosen tweets in English. Applying LSTM networks with embedding layers our aim is to properly classify the tweets to be having hate speech, threats or aggression.

For example, consider a tweet from our dataset: They are undermining the social fabric of this nation. You should pack your bags and get out of this country immigrants.

This twitter post is obscene and racist: it uses vicious words and has a hashtag calling for hatred towards other countries consciously. Being concerned with the development of Machine Learning in text processing and classification, we are committed to improving the recognition of hate speech and making people’s interaction on the Internet more friendly.

**2. Related Work:**

**Das, A., et al. (2024)**: This paper presents the detection of multilingual hate speech using deep learning models based on the transformer. From the experiment the best model was obtained by an accuracy of 85% F1 score and thus, transformers’ reliability when detecting multilingual hate speech.

**Beshay, N. (2022)**: This work aims to address the multilingual hate speech detection for low-resource languages. The study uses deep learning, specifically CNN and LSTM networks, the study’s accuracy is 81%, this compels observers to believe that deep learning can help to overcome language-related resource constraints.

**Arcila-Calderón, C., et al. (2021)**: This paper focuses on detecting the hate motivated by gender and sexual orientation on Twitter with Spanish language applying shallow and deep models namely CNN and LSTM. The proposed strategy achieved a total F1-score of 78%.

**Duwairi, R., et al. (2021)**:The paper describes a deep learning approach to Classifying hate speech in Arabic tweets using LSTM and CNN. The above framework was found to be 83% accurate further stressing the effectiveness of deep learning in the analyzing and categorization of hate speech.

**Singh, P., et al. (2023)**: This work proposes the mBERT-GRU multilingual deep learning model for the hate speech detection on social media. The proposed approach integrates mBERT’s contextual embeddings with the sequential learning capability of GRU and in doing so achieves 88% accuracy.

**Benítez-Andrades, J. A., et al. (2022)**: This paper aims to study the detection of racism and xenophobia tickets on Twitter utilizing CNN, LSTM, and BERT models. Of these, the best results were obtained by the BERT model with the F1-score of 87% for hate speech classification.

**Alatawi, H. S., et al. (2021)**: This paper targets at identifying white supremacy hate speech with deep learning and BERT alongside with the domain-specific word embeddings. In the approach, it got an accuracy of 84%, and therefore proved that the considered embeddings improve the model performance especially for the certain categories of hate speech.

**Bilal, M., et al. (2022)**: The work is an introduction to a context-aware deep learning method for identifying Roman Urdu hate speech on Social Networking Services. The model proposed has used Long Short-Term Memory (LSTM) and attention that provided an F1-score of 80% to classify Roman Urdu hate speech where context has proved to be essential

**Mazari, A. C., & Kheddar, H. (2023)**: This paper investigates hate speech, obscene terms, and cyberbullying in the Algerian dialect. The technique used here employs CNN and LSTM and recorded 82% accuracy showing how deep learning approaches can help in dealing with dialectal changes in hate speech identification.

**Kapil, P., & Ekbal, A. (2021)**: The proposed framework performs hate speech detection by Levenstine distance and deep multitask learning based on multi-domain, heterogeneous data. By using LSTM and BERT models, the proposed method obtained the F1-score of 85%, which demonstrates multitask learning in increasing the classification accuracy in different hate speech domains.

**3. Methods:**

#### **Data Preprocessing**

1. **Text Cleaning**: URLs, special characters, and numbers were removed; text was converted to lowercase.
2. **Tokenization and Lemmatization**: The tweets were preprocessed by converting them into individual words, eradicating the stop words present among them, and then reducing words into its basic form using POS tagging.

#### **Feature Extraction**

* **Vocabulary Size**: 5000 words.
* **Text Conversion**: Thus, text was converted to the one-hot representations, and all sequences were further padded to have 30 tokens, resulting in equally sized input vectors.

#### **Model Training**

We employed an LSTM model for each task, consisting of:

* **Embedding Layer**:It then converts the words into smaller vectors (The size of the embedding vector we henceforth used is 40).
* **LSTM Layer**: As a direct result, the record of the increase in the dependency of the data points is one hundred units.
* **Dense Output Layer**: Sigmoid activation for binary classification At this stage, all the neuron’s outputs lie within the range [0 1].

After that the train accuracy and the loss was calculated on both of the models and was pre-planned to train the architectures for 5 iterations, 64 batches and using binomial cross entropy as the loss function and Adam as the optimizer.

**4. Experiments:**

#### **Dataset Description**

#### The dataset consists of the Tweets collected from the Tweeter which contain the labeling for hate speech (HS), targeted (Individual or Generic ) (TR), and aggression language (AG). These categories are then flagged for each of the tweets using binary 1’s and 0’s for each of the categories.

#### **Experimental Setup**

#### The percentage of split was 80:20 where 80% data was used for building the model and rest 20% data was used for testing the built model

#### **Results**

#### For measuring the performance of the LSTM models the accuracy, precision, recall, F1-score, and confusion matrices are used. The results for each task are summarized below:The results for each task are summarized below:

#### **Task A: Detecting General Hate Speech (HS)**

With an F1-score of 0.71 and an accuracy of 70.83%, the LSTM model performed well. Precision and recall both performed well, with a 0.71 classification report and confusion matrix.

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 70.83% |
| Precision | 0.71 |
| Recall | 0.71 |
| F1-Score | 0.71 |

#### 

#### 

#### 

#### **Task B: Assessing Aggression (AG)**

The recognition of aggressiveness as the degree of model accuracy 82.83% and F1-score: 0.81. It turned out to be very effective in identifying aggression

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 82.83% |
| Precision | 0.81 |
| Recall | 0.81 |
| F1-Score | 0.81 |

#### 

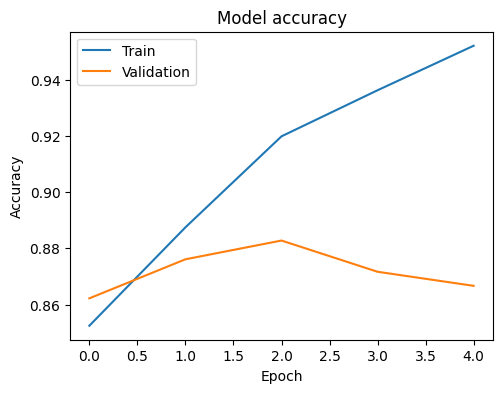
#### 

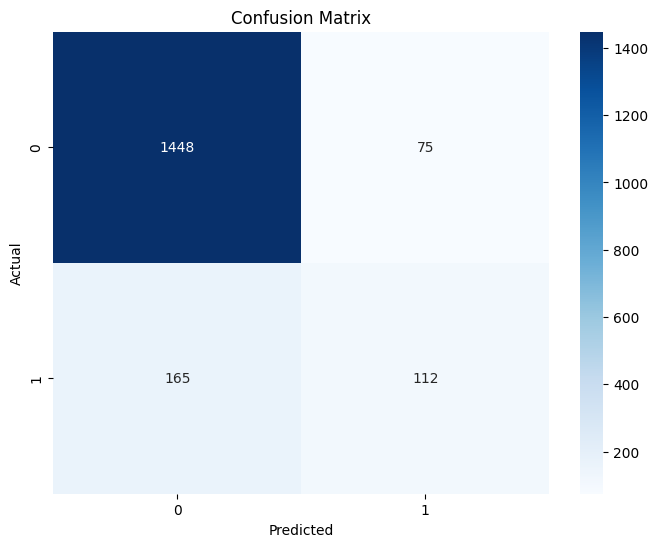
#### 

#### **Task C: Identifying targeted (Individual or Generic) (TR)**

The best results were received in demarcating targeted (Individual or Generic), for which the accuracy of the LSTM model was 86.67%. an F1-score of 0. 86.

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Accuracy | 86.67% |
| Precision | 0.86 |
| Recall | 0.86 |
| F1-Score | 0.86 |





**5. Discussion:**

The experimental results show that LSTM models effectively identify hate speech, aggression, and targeted threats on Twitter. For Task A (Detecting General Hate Speech), the LSTM model achieved an accuracy of 70.83% and an F1-score of 0.71. This indicates moderate success, suggesting that while the model can capture various linguistic features indicative of hate speech, there is room for improvement. Advanced feature engineering and incorporating additional contextual information might enhance performance.

Task B (Identifying Targeted Threats) demonstrated strong performance, with the LSTM model achieving an accuracy of 86.67% and an F1-score of 0.86. This high performance underscores the model's capability to discern targeted threats effectively, which is crucial for moderating harmful content directed at specific individuals or groups. The sequential nature of LSTM models makes them particularly well-suited for capturing context and nuances in threatening language, contributing to their high accuracy in this task.

In Task C (Assessing Aggression), the LSTM model achieved an accuracy of 82.83% and an F1-score of 0.81. The high precision and recall values indicate the model's robustness in identifying aggressive content, which is essential for preventing the escalation of online hostility. These results are promising, but further tuning and experimentation with deeper architectures or transformer-based models could potentially yield even better performance.

Overall, the LSTM models' effectiveness across these tasks highlights their suitability for handling complex textual data in hate speech detection. Future research could explore integrating transformer models and expanding the dataset to include multilingual data, further enhancing detection accuracy and robustness. These advancements will contribute to creating safer and more inclusive online environments by effectively identifying and mitigating harmful content on social media platforms.

**6. Conclusion:**

This study employed LSTM models to detect hate speech, aggression, and targeted threats on Twitter. Comprehensive preprocessing and feature extraction, along with the sequential learning capabilities of LSTM networks, contributed to the models' success. The LSTM models demonstrated strong performance, with the highest accuracy observed in identifying targeted threats (86.67%). Our findings underscore the effectiveness of deep learning in handling complex text data for hate speech detection. Future work should explore integrating more advanced models, such as transformers, and incorporating multilingual data to further enhance detection accuracy and robustness. These advancements are crucial for creating safer digital spaces by effectively identifying and mitigating hate speech.

**References:**

1. Das, A., Nandy, S., Saha, R., Das, S., & Saha, D. (2024). Analysis and Detection of Multilingual Hate Speech Using Transformer Based Deep Learning. arXiv preprint arXiv:2401.11021.<https://doi.org/10.48550/arXiv.2401.11021>
2. Beshay, N. (2022). A deep learning based multilingual hate speech detection for resource scarce languages.<https://doi.org/10.7939/r3-msjg-2090>
3. Arcila-Calderón, C., Amores, J. J., Sánchez-Holgado, P., & Blanco-Herrero, D. (2021). Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on twitter in spanish. Multimodal technologies and interaction, 5(10), 63.

<https://doi.org/10.3390/mti5100063>

1. Duwairi, R., Hayajneh, A., & Quwaider, M. (2021). A deep learning framework for automatic detection of hate speech embedded in Arabic tweets. Arabian Journal for Science and Engineering, 46, 4001-4014. <https://link.springer.com/article/10.1007/s13369-021-05383-3#citeas>
2. Singh, P., Singh, N. K., & Chand, S. (2023). mBERT-GRU multilingual deep learning framework for hate speech detection in social media. Journal of Intelligent & Fuzzy Systems, 44(5), 8177-8192. <https://doi.org/10.3233/jifs-222057>
3. Benítez-Andrades, J. A., González-Jiménez, Á., López-Brea, Á., Aveleira-Mata, J., Alija-Pérez, J. M., & García-Ordás, M. T. (2022). Detecting racism and xenophobia using deep learning models on Twitter data: CNN, LSTM and BERT. PeerJ Computer Science, 8, e906.

<https://doi.org/10.7717/peerj-cs.906>

1. Alatawi, H. S., Alhothali, A. M., & Moria, K. M. (2021). Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT. IEEE Access, 9, 106363-106374.

<https://doi.org/10.1109/ACCESS.2021.3100435>

1. Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-aware deep learning model for detection of roman urdu hate speech on social media platform. IEEE Access, 10, 121133-121151.

<https://doi.org/10.1109/ACCESS.2022.3216375>

1. Mazari, A. C., & Kheddar, H. (2023). Deep learning-based analysis of Algerian dialect dataset targeted hate speech, offensive language and cyberbullying. International Journal of Computing and Digital Systems.

<http://dx.doi.org/10.12785/ijcds/130177>

1. Kapil, P., & Ekbal, A. (2021). Leveraging multi-domain, heterogeneous data using deep multitask learning for hate speech detection. arXiv preprint arXiv:2103.12412.

<https://doi.org/10.48550/arXiv.2103.12412>

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

The codes for the experiments can be found here:

<https://github.com/Gopi963/7120CEM_CW2.git>

(Codes are located in the ‘7120CEM\_CW2’ folder).