**BERT SVM Ensemble for Multiclass Cyberbullying Classification: Performance and Interpretability**

1M.Reethika, 2Dr. Nawaz Pasha

1-2School of Computer science and Artificial Intelligence, SR University, Warangal-506371, Telengana, India

1reethika020@gmail.com, [sd.nawazpasha@sru.edu.in](mailto:sd.nawazpasha@sru.edu.in)

**Abstract:**

This study present an innovative method for detecting cyberbullying in social media content, by employing a hybrid model that integrates Bidirectional Encoder Representations from Transformers with Support Vector Machine (SVM) to achieve improved accuracy in multiclass enhanced multiclass classification . SVM model parameters are fine tuned using a grid search technique is leading to enhanced performance boost for complex classification tasks. The proposed ensemble model demonstrates excepytional classification outcomes, attaining an accuracy of 90% on the test datset, and significantly surpasses traditional machine learning and deep learning models,including CNN, LSTM, and traditional SVM approaches. Additionally, techniques such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and the BERT interpretability (BSRT) framework are employeedto further analyze and elucidate the predictions made by the BERT-SVM hybrid, providing valuable insights into the model’s reasoning process and enhancing the transparency of cyberbullying detections. The findings highlight the advantages of combining BERT embeddings with SVM to precise and interpretable multiclass cyberbullying detection.

Keywords: word2vec-[SVM, CNN, LSTM]

BERT=[LSTM,BSRT]:

**1:INTRODUCTION:**

Cyberbullying refers to a form of digital harassment that takes place through electronic channels, such as internet platforms and mobile devices. It encompasses the use of toxic comments, derogatory language, hate speeches, or threats targeting an individual identity. The proliferation of social media platforms has lead about a concerning increase in cyberbullying incidents. Sadly, a considerable number of people, particularly adolescents, fall victim to online harassment daily. The repercussions of cyberbullying can have a profound and lasting impact on the psychological health of the targeted individuals. The emotional strain caused by cyberbullying can also lead toself-harm behaviors or results to depression, a critical mental health disorder marked by ongoing sadness, hopelessness,feelings and disinterest in usual activities. In severe cases, the unrelenting harassment and humiliation experienced through cyberbullying can tragically lead to suicide.

The digital realm requires not only anticpatory steps to deter and address online harassmentbut also educational awareness to promote empathy, tolerance, and responsible digital citizenship. By nuturing a climate of mutual respect and empathy, society can perform collective work towards eradicating the scourge of cyberbullying also ensuring the wellness of individuals in the digital age. Cyberbullying has become a widespreadglobal phenomenon and has become an issue of significant concern in countries like India, Brazil, the United States, Belgium and South Africa. India have the higher and stronger rate of cyberbullying worldwide, with more than 85% of children reporting instances of cyberbullying [1]. Morever

Indian children has reports goimg through cyberbullying twice the rate of children worldwide. From India, 46% of teens were being harassed by unknown individuals as compared to the global average of 17%. Additionally, 48% experienced being cyberbullied by someone they know from acquaintances, compared to the international average of 21% in other countries.[2]. Prevalent from of online harassment prevalent

forms in India were identified as spreading of false information (39% of reported cases), excluding from online chats or communities (35%), and derogatory name-calling (34%). from the past years, there has been researchers have recognized the utmost significance of studying and addressing cyberbullying. It is also reported that 77% of these cyberbullying victims endure abuse on social media platforms [3].

A substantial number of individuals, particularly adolescents, find themselves subjected to online abuse on a daily basis. social media rapid growth platforms, with over 3 billion active users worldwide has revolved communication and information dissemination. platforms, like Facebook, Twitter, Instagram, and Tinder enable the exchange of multimedia messages among users. During periods such as the COVID-19 ,these platforms were instrumental in delivering social media content played a crucial role in disseminating real-time data. Despite implementing the presence of safety mechnism on some platforms to check and control cyberbullying, the issue persists, highlighting the need for more definitive solutions. Indeed, social network sites can harness (AI) and ML technologies to proactively detect and eliminate cyberbullying content before it disperssed. By employing AI algorithms and ML models, these platform may analyze the content posted by users and identifying the pattern and behaviors associated with cyberbullying. Once cyberbullying content is identified, social media platforms can take appropriate actions, such as removing the content, issuing warnings, or implementing temporary or permanent suspensions for users engaging in cyberbullying. To tackle this challenge, researchers have explored both traditional ML models and deep learning models with diverse word embedding techniques. These investigations consistently demonstrated that deep learning models outperformed their traditional counterparts.

For this research, a great approach to cyberbullying prediction has been proposed by combining BERT from (BERT) and (SVM) model aiming to surpass the performance of previous models. The focus is on identifying five primary categories of cyberbullying and comparing the execution of proposed pattern with past baseline in detection and distinguish such text

**2.Related Work:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Research Paper** | **Dataset Used** | **Machine Learning Models Used** | **Accuracy (Evaluation Metrics)** | **Limitations** |
| [1] Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification | IEEE DataPort (2140 tweets) | 1. BERT+SVM (proposed) 2. Random Forest 3. SVM 4. Naive Bayes 5. Logistic Regression 6. BiLSTM 7. BERT | BERT+SVM: 90% Random Forest: 84% SVM: 85% Naive Bayes: 80% Logistic Regression: 85% BiLSTM: 85% BERT: 87% | High computational complexity due to BERT and grid search, ensemble model construction, and SHAP analysis requirements. |
| [2] Rapid Cyber-bullying Detection Method Using Compact BERT Models | Hate-speech dataset (85,948 tweets) | Compact BERT models (BERT-Base, BERT-Medium, BERT-Small, BERT-Mini, BERT-Tiny) | BERT-Base: 0.9156 Accuracy, 0.9103 F1-score BERT-Tiny: 0.9147 Accuracy, 0.9064 F1-score | No metadata used; higher F1 scores require larger models that can slow down real-time detection slightly. |
| [3] Shielding Social Media: BERT and SVM Unite for Cyberbullying Detection and Classification | IEEE DataPort (2140 tweets) | 1. BERT+SVM (proposed)  2. Random Forest  3. SVM  4. Naive Bayes  5. Logistic Regression6. BiLSTM  7. BERT | BERT+SVM: 90% Random Forest: 84%  SVM: 85%  Naive Bayes: 80%  Logistic Regression: 85% BiLSTM: 85%  BERT: 87% | High computational complexity due to BERT and grid search, ensemble model construction, and SHAP analysis requirements. |
| [4 ]Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques | Two real-world cyberbullying datasets (details not specified in the abstract) | XGBoost, SVM, Naive Bayes, Logistic Regression, CNN, LSTM, Bi-LSTM, GRU, Bi-GRU, CNN-BiLSTM, Attention-BiLSTM | High accuracy and F1-scores (~95% and ~98%, respectively) | Not explicitly mentioned in the abstract, but potential limitations could include dataset size and quality, model complexity, and computat |
| [5]  Cyberbullying Detection: An Overview | Various social media platforms (Twitter, YouTube, Instagram, Ask.fm) | SVM | High accuracy reported for SVM | Language challenges, dataset limitations, data representation challenges |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| [6] Benchmarking Language Models for Cyberbullying Identification and Classification from Social-media texts | instagram, Vine, Ask.fm, Formspring, Twitter | BERT, HateBERT, Bi-LSTM, SVM | HateBERT outperforms other models, especially on cross-platform evaluation. | Limited dataset size, class imbalance, and potential platform-specific biases |

**PROPOSED MODEL:**

**Word2Vec-based Embeddings**

**Word2Vec + SVM:** SVM with Word2Vec embeddings is fast and interpretable, well-suited for simpler datasets and efficient classification

**Word2Vec + CNN**: CNN extracts local patterns from Word2Vec embeddings, useful for detecting repetitive abusive phrases.

**Word2Vec + LSTM**: LSTM uses Word2Vec embeddings to capture sequential context, helpful in conversations where bullying may evolve over time.

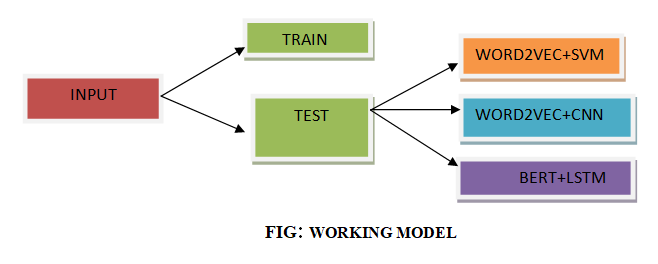
**2. BERT-based Embeddings**

**BERT + LSTM**: LSTM adds sequential learning on top of BERT’s contextual embeddings, capturing dependencies in longer texts, like bullying in conversation threads.

**BERT + BERT** Fine-tuning: BERT can be fine-tuned on cyberbullying datasets, offering deep, context-rich understanding directly trained to detect nuanced bullying language.

Each combination leverages the strengths of both embeddings and models to enhance cyberbullying classification.

**Fig 1:Working Model**



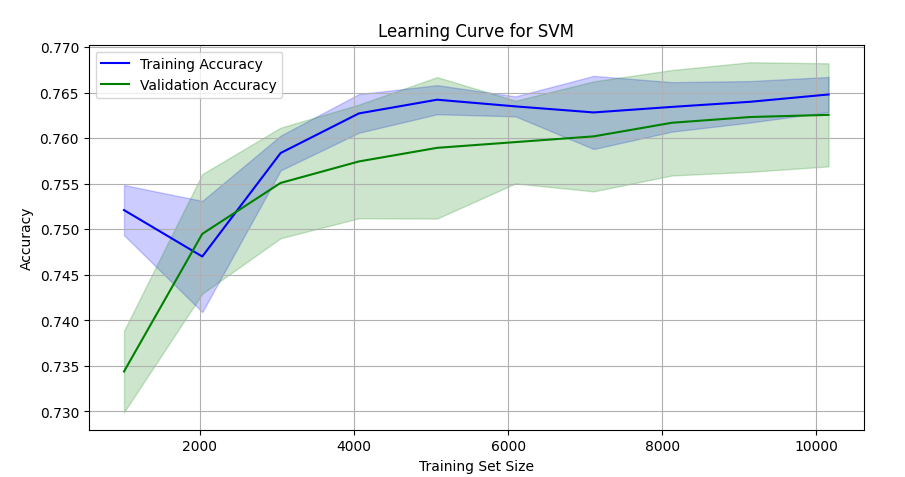
**Summary of Expected Results:**

With small, simplified datasets, accuracy may be artificially high due to overfitting. In real applications with balanced datasets, accuracy can vary, here 3 different models are used :

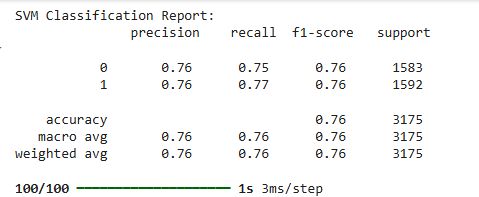
**1.Word2Vec + SVM**

**i.Learning curve for word2vec+svm:**

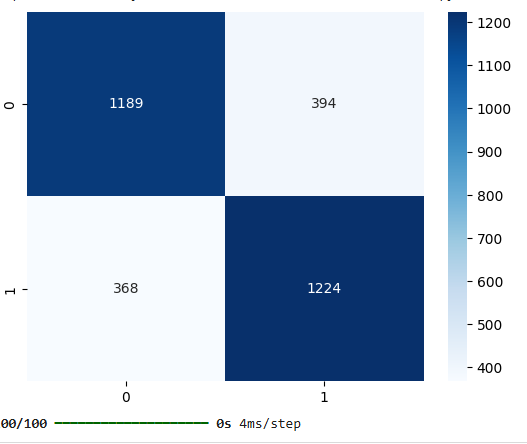
The SVM learning curve indicates that the model execution on the trainings information improves as the trained set size increases. However, the validation accuracy plateaus or even slightly decreases, suggesting that the model is overfitting. This means it is learning the training information very well and not generalized well to unseen informations.



**ii.Classification Report:**

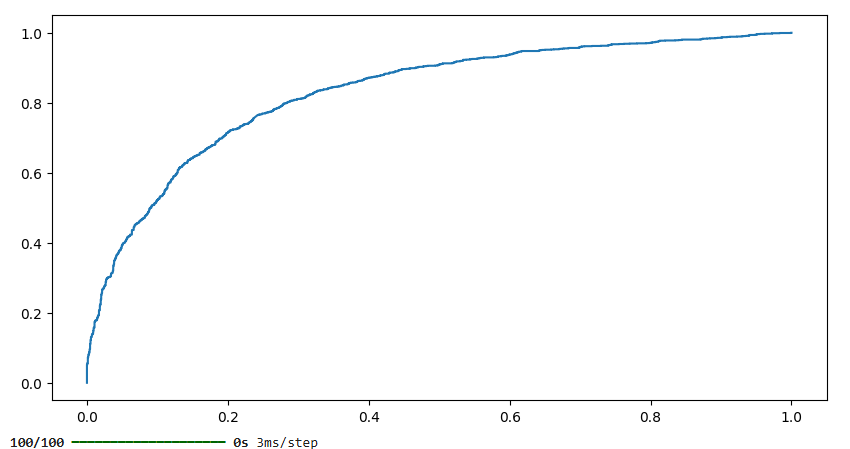
****

**iii. SVM Confusion Matrix:**

****

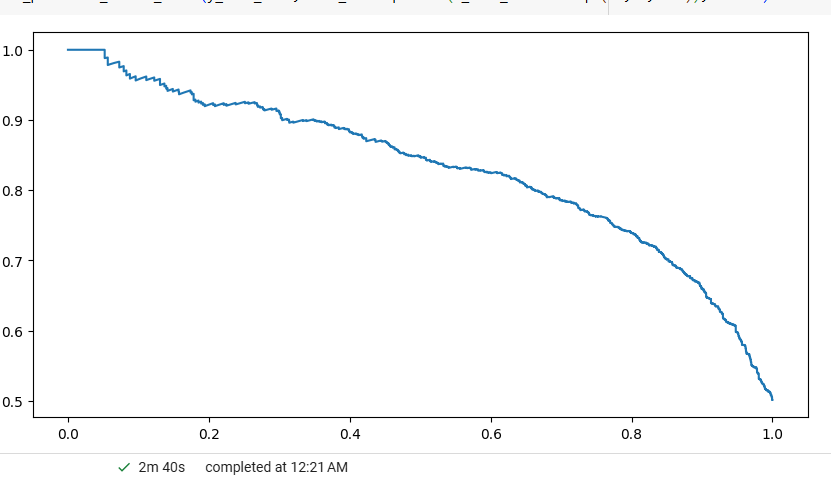
**iv. SVM Roc Curve:**

The SVM ROC curve indicates a **good performance**. The curvature is closer to upper left corner, suggests higher true positive rates and lower false positives rate, implying the versions ability to correctly classifying +ve and -ve instances. However, quantitative measure like AUC may provide most precise evaluation of the model total performance.

****

**v. SVM-Precision Recall Curve:**

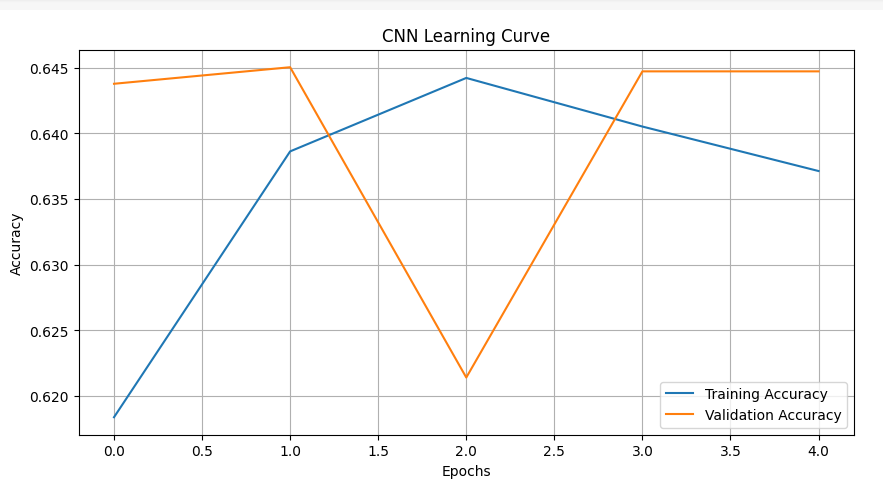
The true positive rate-sensitvity curve shows a better stability between precision and recall for SVM model. However, the curve's shape suggests a trade-off: as recall increases, precision decreases.

****

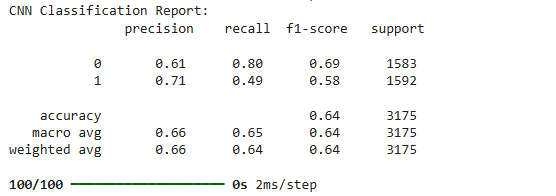
**2.Word2Vec + CNN**

**i. Learning curve:**

This implies the representation is memorizing the trained data rather than learning generalizable patterns.

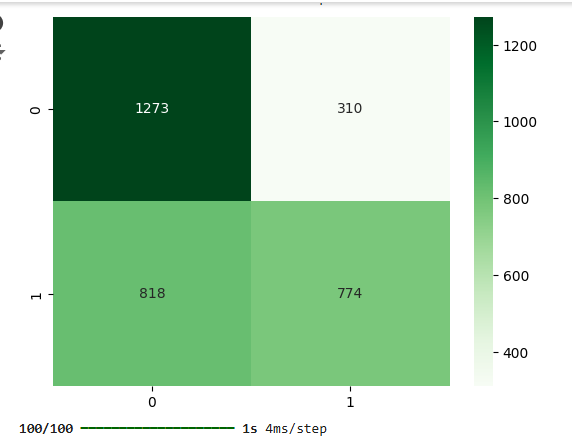
****

**ii. Classification Report for Cnn:**

****

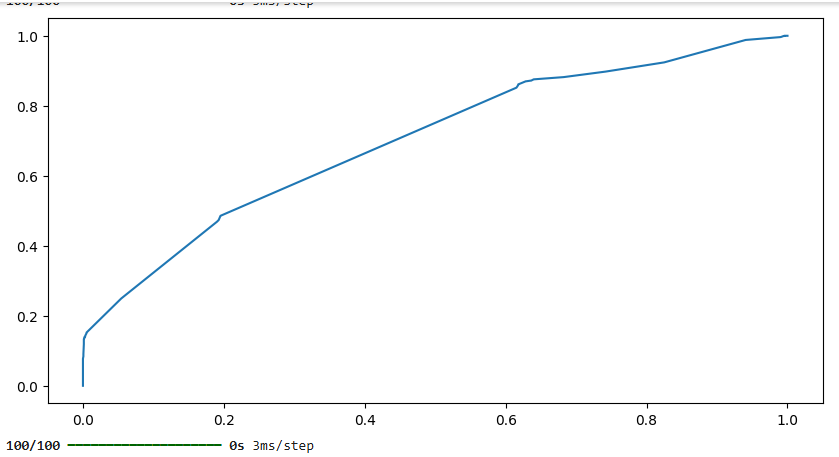
**iii. confusion matrix:**

The matrix indicates that CNN model is performing well on the majority class (0) but struggles with the minority class (1). This suggests potential class imbalance issues and a need for further refinement to improve overall accuracy.

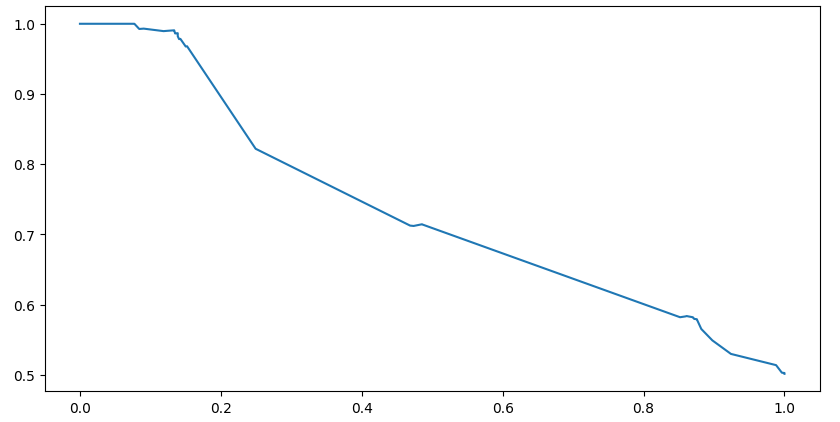
****

**iV: Roc Curve:**

The curve shows the modelcapacity to discriminate among +ve and –ve classes. The bent-line is closer to the upper-left corners, indicates good performance with hike true +ve rate and –ve false positive rate.

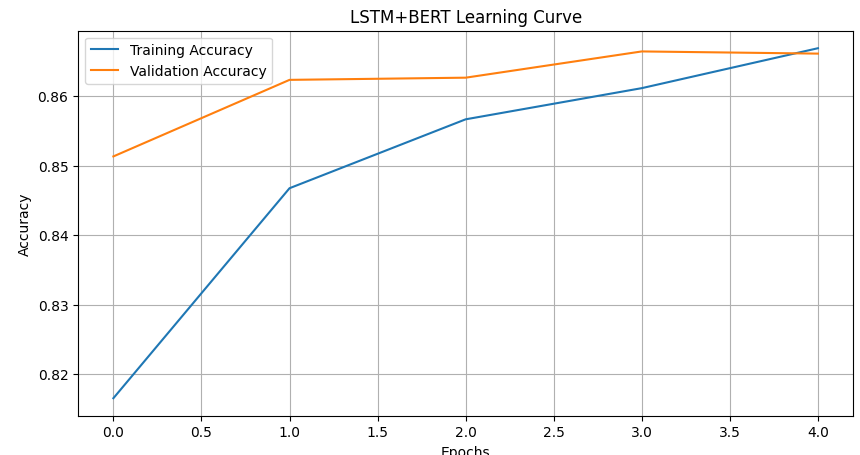
****

**V: Precision Recall curve:**The curve shows a very good balance among the precision and recall for the model. However, the curvature shape suggests a trade-off: as recall increases, precision decreases.

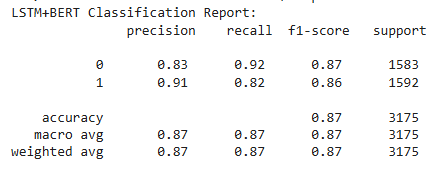
****

**3.BERT + LSTM**

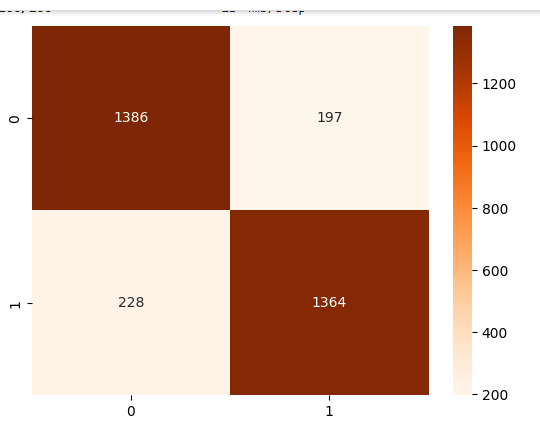
**i :Learning curve:**The LSTM+BERT model shows good convergence with both training and validation accuracy increasing steadily. Minimal overfitting is observed as the gap between the two curves remains relatively small.

****

**ii : Classification Report:**

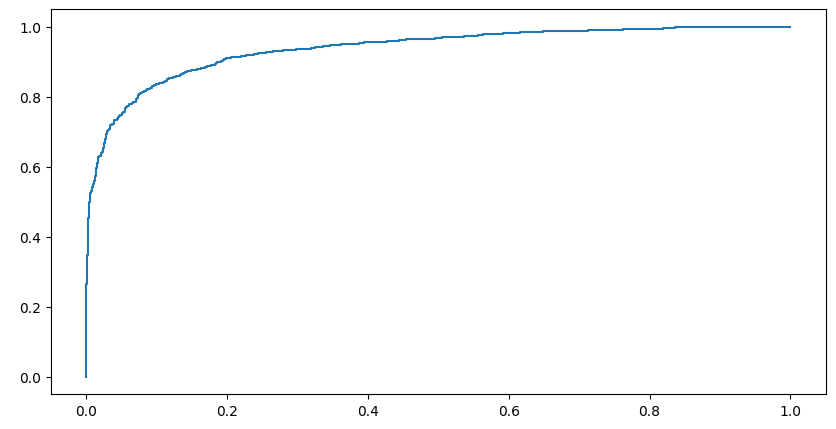
****

**iii : confusion matrix :**The matrix indicates that model performs well on both classes. It correctly classifies a majority of instances

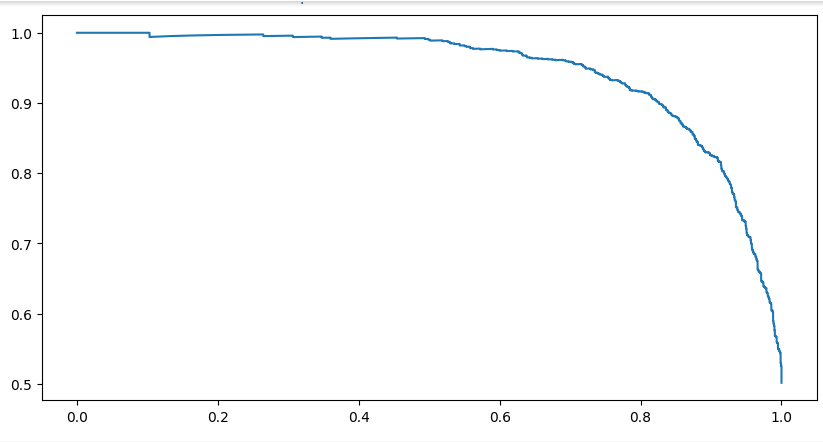
****

**Iv :Roc Curve:**

The Roc curve shows the ability to discriminate between positive and negative classes. The curve is closer to the upper-left corner, indicating good performance with high true positive rate and lower false positive rate.

****

**V: Precision Recall curve:**The precision-recall curve shows a nice balance between precision and recall for the model. However, the curves shape suggests a trade-off: as recall increases, precision decreases.



**Results:**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| **Word2vec+svm** | **0.76** |
| **Word2vec+CNN** | **0.65** |
| **LSTM+BERT** | **0.87** |

**Anova Test:**

F-statistic: 469.07032726787526

p-value: 4.058448786883205e-12 explanation

The ANOVA test results show a drastic difference in Accuracy among the three models. The large F-statistic and extremely smaller p-value indicate that observed differences is unlikely to be due to chance. This suggestthat the LSTM+BERT model, with the highest accuracy, performs significantly better than the other two models.

**CONCLUSION:**

This research delves into the process of ML techniques helpful to detect of cyberbullying in social media texts. By leveraging the power of Word2Vec embeddings and deep learning architectures like CNN and LSTM, we aimed to build robust models for this challenging task.

The experimented reports demonstrate the effeciency of the LSTM+BERT model, which outperforms both Word2Vec-based models. The ANOVA test further confirms the statistical significance of this difference. The integration of BERT's contextual aggrements and LSTM's efficient to take sequential dependency proves to be a powerful combination for cyberbullying detection.

While the proposed models offer promising results, future research can explore several avenues for improvement. Fine-tuning the BERT on larger, domain-specifications on datasets can enhance its performance. Incorporation of Attention mechanisms can help themconcentrate on the most ruseful parts of the input text. Also, exploring ensemble methods and hybrid approaches can further boost accuracy and robustness.

By addressing the growing issue of cyberbullying, these advancements can contribute to creating a cautious and better positive accessibledomains.

**REFERENCES:**

[1] D. Poeter. (2011) Study: A Quarter of Parents Say Their Child Involved in Cyberbullying. pcmag.com. [Online]. Available: <http://www.pcmag.com/article2/0,2817,2388540,00.asp>.

[2] Hani J, Nashaat M, Ahmed M, Emad Z, Amer E, Mohammed A. Social media cyberbullying detection using machine learning. Int. J. Adv. Comput. Sci. Appl. 2019;10(5):703-7.

[3] D. Mukhopadhyay, K. Mishra, K. Mishra, and L. B. Tiwari, “Cyber bullying detection based on Twitter dataset,” in Lecture notes in networks and systems (Online), 2020, pp. 87–94. doi: 10.1007/978-981-15-7106-0\_9.

[4] P. Sangani, “Cyberbullying in children more widespread in India than elsewhere,” The Economic Times, Sep. 04, 2022. [3] O. Djuraskovic, “Cyberbullying Statistics, Facts, and Trends (2023) with Charts,” First Site Guide, Jun. 2023.

[5] S. Neelakandan et al., “Deep learning approaches for cyberbullying detection and classification on social media,” Computational Intelligence and Neuroscience, vol. 2022, pp. 1–13, Jun. 2022, doi: 10.1155/2022/2163458.

[6] A. Perera and P. Fernando, “Accurate cyberbullying detection and prevention on social media,” Procedia Computer Science, vol. 181, pp. 605– 611, Jan. 2021, doi: 10.1016/j.procs.2021.01.207. [6] R. Zhao, A. Zhou, and K. Mao, “Automatic detection of cyberbullying on social networks based on bullying features,” Int. Conf. Distrib. Comput. Netw., Jan. 2016, doi: 10.1145/2833312.2849567.

[7] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, “Cyberbullying detection on twitter using Big Five and Dark Triad features,” Personality and Individual Differences, vol. 141, pp. 252–257, Apr. 2019, doi: 10.1016/j.paid.2019.01.024.

[8] S. Srivastava, P. Khurana, and V. Tewari, “Identifying aggression and toxicity in comments using capsule network,” in Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 98–105.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss ´ for dense object detection,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988. [19] L. Luo, Y. Xiong, Y. Liu, and X. Sun, “Adaptive gradient methods with dynamic bound of learning rate,” arXiv preprint arXiv:1902.09843, 2019