IRSTI 28.23.37

10.51889/2959-5894.2024.87.3.010

Zh. Zh. Azhibekova[1], A.O. Aliyeva[2], N.F. Sarsenbiyeva[2], B.S. Kaldarova[2], A.B.Toktarova[3*]

[1]Asfendiyarov Kazakh National Medical University, Almaty, Kazakhstan
[2]South Kazakhstan Pedagogical University named after Ozbekali Zhanibekov, Shymkent, Kazakhstan
[3]M. Auezov South Kazakhstan University, Shymkent, Kazakhstan
[*]*e-mail: toktar.aigerim@list.tu*

BIDIRECTIONAL LONG SHORT-TERM MEMORY IN HATE SPEECH DETECTION PROBLEM ON NETWORKS

Abstract

The pervasive problem of hate speech on social media has received much attention in the fields of computational linguistics and artificial intelligence. The abstract summarizes the results of a groundbreaking study investigating the use of Bi-LSTM models to detect and analyze hate speech. This study highlights the need for modern machine learning algorithms to analyze the large and ever-changing volume of material on social media. The main idea is to accurately detect and reduce instances of hate speech. The paper reviews several procedures used to detect hate speech and traces their evolution over time. It highlights the shortcomings of traditional models and emphasizes the need for more sophisticated, context-aware approaches. The use of Bi-LSTM structures, known for their effectiveness in capturing long-term relationships in sequential data, marks a methodological advance in the field. The results of this study will demonstrate that Bi-LSTM models have a greater ability to understand the complexities of language on social media. The results of research we offer a more efficient and effective method for detecting hate speech. This study makes a valuable contribution to the ongoing debate on digital politeness through rigorous experimentation and analysis. It proposes a robust framework that will adapt and extend across various social media platforms to create safer online communities.

Keywords: BiLSTM, LSTM, AI, NLP, social media.

Ж.Ж. Ажибекова[1],А.О. Алиева[2], Н.Ф. Сарсенбиева[2], Б.С. Қалдарова[2], А.Б. Тоқтарова[3]
[1]С. Ж. Асфендияров атындағы Қазақ ұлттық медицина университет, Алматы қ., Қазақстан
[2]Өзбекәлі Жәнібеков атындағы Оңтүстік Қазақстан педагогикалық университет, Шымкент қ., Қазақстан
[3]М.Ауезов атындағы Оңтүстік Қазақстан Университеті, Шымкент қ., Қазақстан

ЖЕЛІДЕГІ БЕЙӘДЕП СӨЗДЕРДІ АНЫҚТАУДА ЕКІ ЖАҚТЫ ҰЗАҚ МЕРЗІМДІ ЖАДТЫ ҚОЛДАНУ

Аңдатпа

Есептеу лингвистикасы мен жасанды интеллект әлеуметтік желілердегі бейәдеп сөздердің күн санап өсу мәселесіне назар аударуда. Ұсынылып отырған ғылыми зерттеу жұмысында Bi-LSTM үлгілері арқылы бейәдеп сөздерді анықтау және талдау бойынша жаңашыл зерттеулерді ұсынады. Бұл зерттеу әлеуметтік медиадағы үнемі өзгеретін деректер қорының үлкен көлемін талдау үшін заманауи машиналық оқыту алгоритмдерінің маңыздылығын көрсетеді. Ғылыми зерттеу жұмысының мақсаты – бейәдеп сөздер риторикасы жағдайларын анықтау және азайту. Мақалада бейәдеп сөздерді анықтаудың әртүрлі әдістері және олардың уақыт өте келе қалай өзгергені қарастырылады. Бұл дәстүрлі анықтау әдістерінің кемшіліктерін көрсетеді және контекстті анық тани алатын неғұрлым тиімді тәсілдердің қажеттілігін көрсетеді. Біріктірілген деректерде ұзақ мерзімді жадты пайдалануда тиімділігімен белгілі Bi-LSTM құрылымдарын пайдалану осы саладағы әдістемелік серпіліс болып табылады. Зерттеу нәтижелері Bi-LSTM үлгілері әлеуметтік медиа тілінің күрделілігін жақсы түсінетінін көрсетті. Осылайша, ұсынылып отырған ғылыми зерттеу жұмысыбейәдеп сөзді риториканы анықтаудың тиімді әдісін ұсынады. Эксперименттер мен талдаулар арқылы бұл зерттеу

цифрлық «сыпайылық» туралы еңбектерге қосымша қарқын бере алады. Түрлі әлеуметтік медиа платформаларында өзгертуге және кеңейтуге болатын сенімді құрылым онлайн қауымдастықтардың қауіпсіздігін қамтамасыз етеді деп күтілуде.

Түйін сөздер: BiLSTM, LSTM, AI, NLP, әлеуметтік желі.

Ж.Ж. Ажибекова[1],А.О. Алиева[2], Н.Ф. Сарсенбиева[2], Б.С. Қалдарова[2], А.Б. Тоқтарова[3]
[1]Казахский Национальный медицинский университет имени С. Д. Асфендиярова, г.Алматы, Казахстан
[2]Южно-Казахстанский педагогический университет имени Өзбекәлі Жәнібеков, г.Шымкент, Казахстан
[3]М.Ауезов Южно-Казахстанский университет, г.Шымкент, Казахстан

ДВУСТОРОННЯЯ ДОЛГОСРОЧНАЯ ПАМЯТЬ В ПРОБЛЕМЕ ОБНАРУЖЕНИЯ РЕЧИ НЕНАВИСТИ В СЕТЯХ

Аннотация

Компьютерная лингвистика и искусственный интеллект привлекли внимание к растущей проблеме речи ненависти в социальных сетях. В исследовательской статье представлены новаторские исследования по обнаружению и анализу ненавистныхречей с использованием моделей Bi-LSTM. Это исследование подчеркивает важность современных алгоритмов машинного обучения для анализа больших объемов постоянно меняющихся данных социальных сетей. Цель состоит в том, чтобы выявить и сократить случаи ненавистнической риторики. В статье рассматриваются различные методы определения речи ненависти и то, как они менялись с течением времени. Это подчеркивает недостатки традиционных методов и подчеркивает необходимость более сложных подходов, которые могут четко распознавать контекст. Использование фреймворков Bi-LSTM, известных своей эффективностью при использовании долговременной памяти в объединенных данных, является методологическим прорывом в этой области. Результаты исследований показывают, что модели Bi-LSTM могут лучше понимать сложность языка социальных сетей. Таким образом, они обеспечивают эффективный способ обнаружения речи ненависти. Благодаря экспериментам и анализу это исследование может придать импульс работе над цифровой «вежливостью». Надежная структура, которую можно модифицировать и расширять на различных платформах социальных сетей, обеспечивает безопасность онлайн-сообществ.

Ключевые слова: BiLSTM, LSTM, AI, NLP, социальные сети.

Main provisions

The utilization of Bi-LSTM structures, renowned for their efficacy in capturing enduring connections in sequential data, signifies a methodological progression in the sector. The findings of this study will illustrate that Bi-LSTM models has a superior capacity to comprehend the intricacies of language on social media. Our research findings present a superior and more proficient approach to identify hate speech. This paper provides a vital contribution to the continuing discussion on digital civility by conducting thorough experiments and analysis.

Introduction

In the evolving digital communication landscape, social media has become a place for active participation and extensive exchange of ideas. This degree of transparency has also facilitated the widespread dissemination of hate speech, a pernicious trend that threatens the integrity and inclusiveness of online groups. In response to the urgent need to address this problem, numerous computational methods were developed that skillfully detect and mitigate hate speech. The use of Bi-LSTM networks has proven to be a promising area of research. Bi-LSTM networks, due to their ability to evaluate sequential data in both forward and backward directions, offer a deep understanding of language context, making them particularly adept at tackling the intricacies of hate speech identification [1].

Recent research has highlighted the difficulties in automatically detecting hate speech, including the complex structure of language and the dynamic characteristics of online communication. Traditional machine learning models have struggled to capture these intricacies, often ignoring the

contextual connections needed to distinguish between hate speech and benign communication. In contrast, deep learning methods, particularly Bi-LSTM models, have shown significant promise by leveraging their ability to understand long-term correlations in text [1]. Using Bi-LSTM networks to detect hate speech on social media is a significant advance in the fight against online toxicity. Bi-LSTM networks can identify hate speech patterns with greater accuracy than single-track models because they take into account the linguistic context of both preceding and subsequent elements. Moreover, combining these models with social media data facilitates continuous surveillance and immediate intervention, offering a proactive approach to maintaining digital civility. However, using Bi-LSTM models in real-world settings requires addressing several challenges, including the need for sufficient training data and the computational demands associated with processing large datasets.

Moreover, the ethical implications of automated filtering of information on social media raise concerns about censorship and potential restrictions on free speech. Despite these limitations, the development of sophisticated machine learning models such as Bi-LSTM represents a significant opportunity for exploration and use in the ongoing work to mitigate online hate speech [4]. Using Bi-LSTM networks to detect hate speech on social media holds significant promise for improving the safety and inclusiveness of online platforms. This work aims to evaluate the effectiveness of Bi-LSTM models in detecting hate speech by leveraging recent advances in machine learning and natural language processing to address this pressing issue. This study significantly advances the emerging field of computational linguistics and its use in social media moderation. It achieves this by carefully examining current literature and incorporating new findings [5].

Related works

There has been extensive research on the detection and mitigation of hate speech in online environments, with numerous studies exploring different methods and techniques. This section provides a brief overview of the relevant literature, intending to situate the current research within the broader framework of hate speech detection.

Many studies have explored the use of traditional machine learning methods for hate speech detection. For example, [6] used a support vector machine (SVM) to categorize hate speech, highlighting the importance of feature engineering to achieve accurate results. Similarly, [7] used random forests to detect hate speech, illustrating the effectiveness of ensemble methods in solving this challenging problem.

Moreover, deep learning models have become a critical component in hate speech identification research. Recurrent neural networks (RNNs), as outlined in [8], are used to capture sequential relationships in text data. They have demonstrated the ability to detect hate speech patterns. Convolutional neural networks (CNNs), as explained in [9], are used to extract meaningful features from text input, leading to improved accuracy in hate speech detection.

The study in [8] incorporated self-attention mechanisms into LSTM networks, effectively capturing salient subtleties in hate speech content. Transformer models, as mentioned in [7], are also used for this purpose, leveraging their ability to handle long-term dependencies and contextual information.

Researchers have explored ensemble methodologies that go beyond the study of individual machine learning systems. The study in [10] used a combination of different models, including LSTMs and CNNs, to achieve meaningful results in hate speech detection. Ensemble approaches highlight the ability to leverage the strengths of multiple algorithms through their combination.

Another area of research has focused on the integration of specific information and linguistic features. Integrating linguistic attributes into their hate speech detection algorithm, illustrating the importance of linguistic context in identifying hate speech instances. In addition, [11] presented a graph-based approach to detect hate speech by learning semantic relationships between words.

Moreover, transfer learning has gained significant popularity in this area. 10 Demonstrating the transferability of pre-trained models by using extended language models for this specific purpose.

Research on hate speech detection covers a variety of methods. The literature review explains the different strategies and approaches used to tackle the challenging problem of detecting hate speech on online platforms. This study aims to improve the current discourse by evaluating the effectiveness of BiLSTM networks in this important area.

Research methodology

BiLSTM is an advanced variant of recurrent neural networks (RNNs) designed to capture sequential dependencies in data by simultaneously analyzing inputs from both previous and subsequent time steps. The proposed study highlights the numerous advantages of BiLSTM, making it very suitable for the stated purpose. Figure 1 depicts the block diagram of the BiLSTM network.

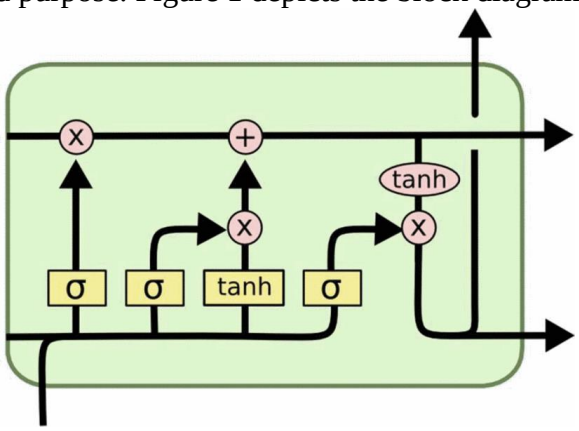

Figure 1. Bidirectional Long Short-Term Memory (BiLSTM) Network

A significant advantage of BiLSTM is its ability to efficiently express and capture distant relationships in text data. Conventional one-way RNNs process text in a single direction, limiting their ability to capture contextual information dispersed within a sentence or text segment. BiLSTM mitigates this limitation by using two hidden layers, where one layer processes the input sequence from left to right and the other from right to left. This allows the network to understand dependencies that span the entire sequence.

BiLSTM networks show significant potential in improving hate speech detection on social media platforms. Their ability to emulate broad links and handle sequences of varying lengths makes them uniquely suited to the dynamic and complex nature of online text data.

Figure 2 depicts the mesh architecture, primarily characterized by the presence of gates in the LSTM and BiLSTM frameworks.

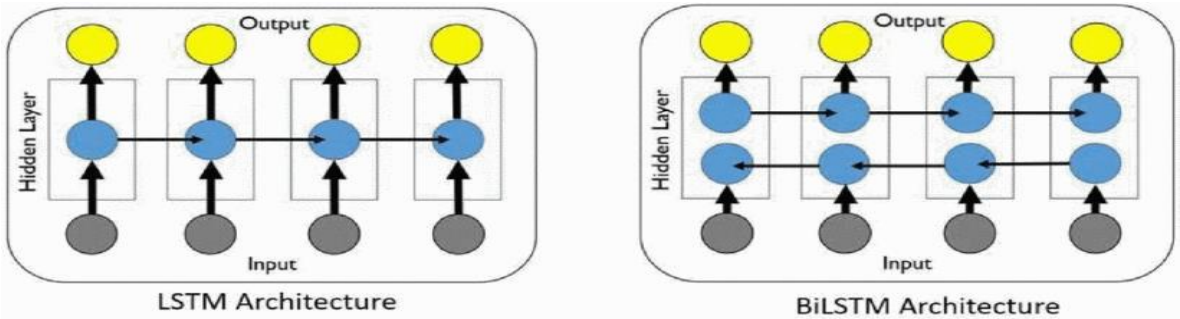

Figure 2. Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) networks

Equation (1) illustrates examples of computational approaches relevant to these specific gate kinds.

$$input(t) = \sigma(W_i x(t) + V_i h(t-1) + b_i) \quad (1)$$

Equation (2) explains the computational method underlying the operation of the forget gate inside a cell. In the equation, Wf and Vf denote the weights associated with the forget gate, which are critical to identifying the data in the cell that requires deletion. It can be concluded that Wf and Vf act as weight parameters of the forget gate.

$$forget(t) = \sigma\left(W_f x(t) + V_f h(t-1) + b_f\right) \tag{2}$$

Equation (1) explains the computational procedure performed by the input gate in the cellular structure. In this equation, h(t-1) denotes the output from the previous cell, x(t) refers to the input in the current cell, and σ denotes the sigmoid function.

$$\tilde{C}(t) = \tanh\left(W_c x(t) + V_c h(t-1) + b_c\right) \tag{3}$$

$$C(t) = forget(t) \cdot C(t-1) + input(t) \cdot \tilde{C}(t) \tag{4}$$

The update mechanisms are defined by equations (3) and (4) as follows: Equation (3) describes a candidate memory block that is tasked with generating alternative update data, while Equation (4) defines the procedure for updating the cell status. The updated data is then integrated with information from the forget gate, resulting in the creation of a new state. In this scenario, Wc and Vc represent weight parameters that dictate the new state.

$$output(t) = \sigma\left(W_0 x(t) + V_0 h(t-1) + b_0\right) \tag{5}$$

$$h(t) = output(t) \cdot \tanh\left(C(t)\right) \tag{6}$$

The calculation method for determining the output gate is explained by equations (5) and (6). In the initial step, a sigmoid layer is used to determine the activation status of the cell. The next step involves applying the hyperbolic tangent (tanh) function to the revised cell status. The last step involves multiplying the current cell state by the output gate state at time t, which yields the output represented as h(t). Vo denotes the weight parameter associated with the output gate. The cell is an integral part of the LSTM neural network design functionality, and understanding it is vital to evaluate the performance of the framework. The fundamental structure serves as a key basis for the development of a bidirectional LSTM (BiLSTM) network, which is designed to extract critical features from data. The regular LSTM framework outperforms its bidirectional counterpart in its ability to capture contextual information [12]. The bidirectional LSTM network utilizes data from both previous and subsequent timestamps by integrating forward and backward time series, improving the accuracy of time series predictions.

Evaluation Criteria

To accurately evaluate the performance of social media hate speech detection models, comprehensive evaluation criteria must be used. For this study, we use a set of well-established evaluation criteria.

Equation (7) illustrates the formula for the accuracy evaluation parameter [13].

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{7}$$

Equation (8) demonstrates formula of precision evaluation parameter:

$$precision = \frac{TP}{TP + FP} \tag{8}$$

Equation (9) demonstrates formula of recall evaluation parameter:

$$recall = \frac{TP}{TP + FN} \tag{9}$$

Equation (10) demonstrates formula of F-score evaluation parameter:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{10}$$

In this study, these evaluation parameters are used to conduct a comprehensive evaluation of the BiLSTM model's performance in detecting hate speech on social media. These measures collectively provide insight into the model's ability to distinguish hate speech from non-hate speech instances while reducing both false positives and false negatives, thereby contributing to a more efficient and detailed evaluation.

Results of the study

This section presents the results obtained using the BiLSTM network for profanity identification. Figure 3 shows the confusion matrix obtained when classifying the text into seven distinct groups. The results of this study confirm that the BiLSTM network is suitable and effective for classifying offensive language. The obtained results demonstrate the competence of the model in accurately detecting offensive language in a given text, highlighting its ability to effectively classify text data into multiple categories. This verification highlights the effectiveness of the BiLSTM network as an important asset in solving problems related to profanity detection. It strengthens control over online content and contributes to a safer digital environment.

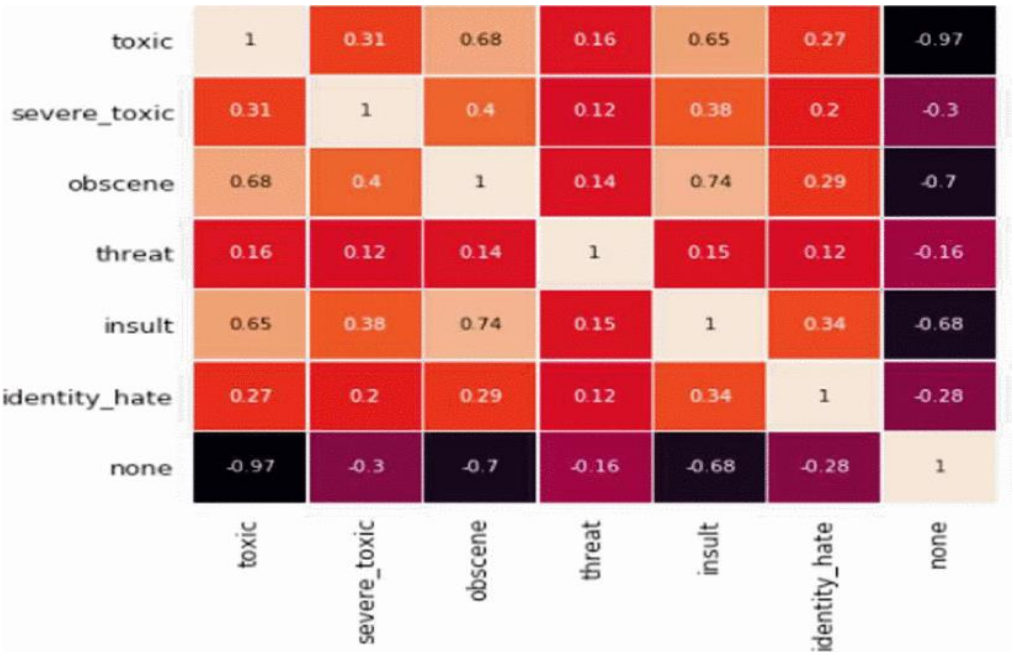

Figure 3. Confusion matrix for the classification of five kinds

Figure 4 illustrates the results of the confusion matrices. This study aims to evaluate the relative performance of different machine learning algorithms in differentiating hostile language from positive and neutral statements by analyzing their confusion matrices. These results improve our understanding of the strengths and weaknesses of each methodology and can help in choosing the most appropriate strategy for detecting objectionable language in this multi-class context.

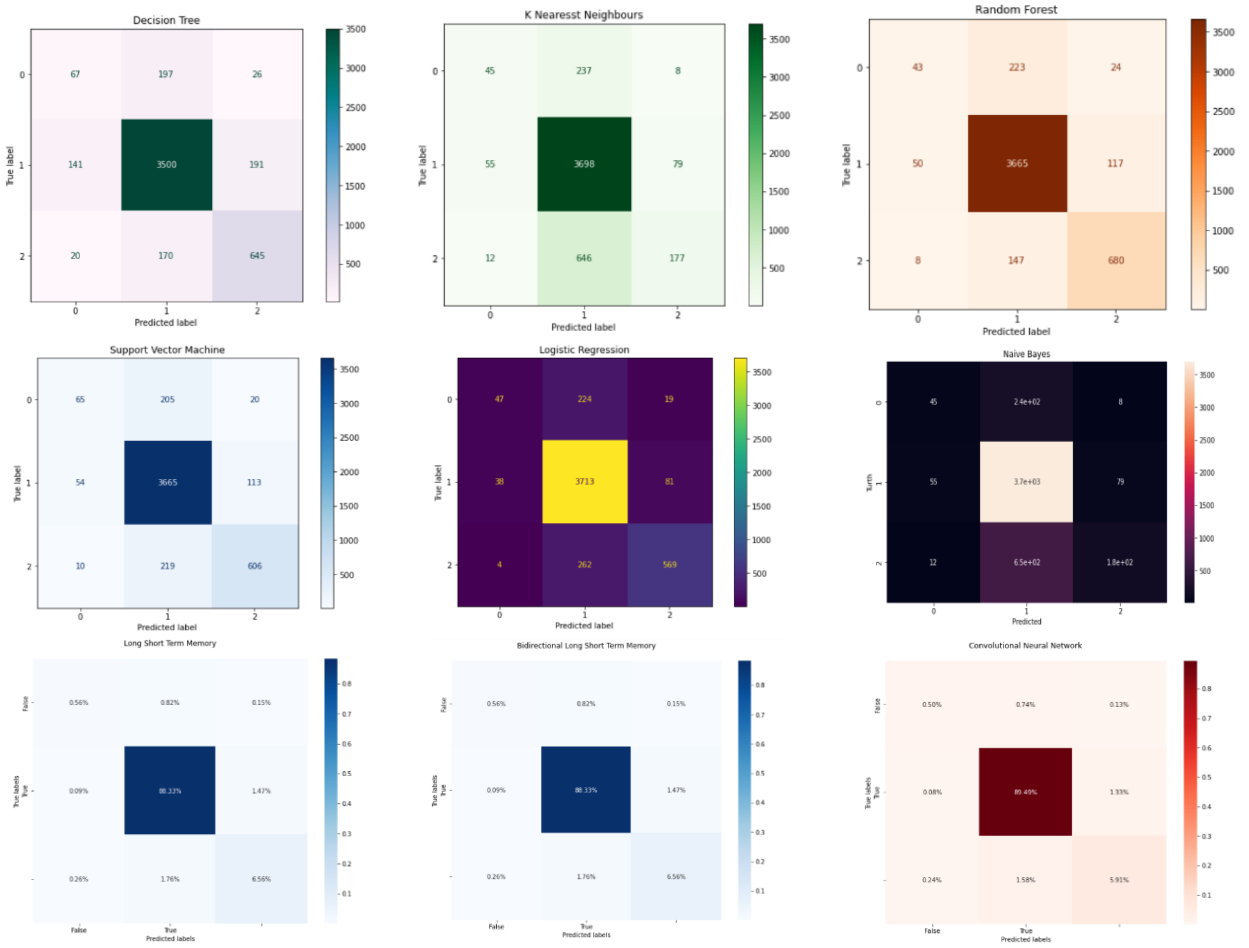

Figure 4. The utilization of confusion matrices yields outcomes in the identification of hate speech.

Figure 5 presents a comprehensive comparison of several machine learning algorithms, with a special focus on the well-studied BiLSTM network and its AUC-ROC performance.

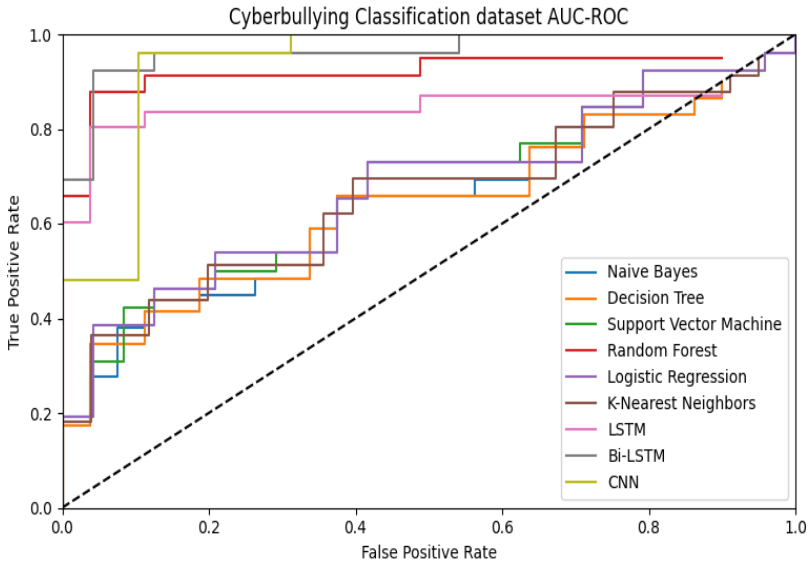

Figure 5. Results of the AUC-ROC curve

This evaluation focuses exclusively on the binary classification issue of offensive language recognition. The graphical representation of AUC-ROC curves facilitates a comprehensive assessment of relative performance, helping to identify the most effective algorithmic approach for the main task of offensive language classification.

The AUC-ROC curves clearly illustrate the balance between the performance of correctly identifying offensive language and incorrectly identifying non-offensive speech, thereby providing valuable insights into the discriminatory capabilities of the models. Importantly, the results show that the BiLSTM network has a remarkable advantage in producing exceptional performance even at the initial stages of the training process. The results show that the investigated BiLSTM network exhibits fast learning and high discriminatory ability to classify unwanted language into two categories. The initial success in achieving exceptional results highlights the potential of this tool as an effective means of recognizing unwanted language, thereby contributing to improved content moderation and a safer online environment.

Discussion

In discussion, improving hate speech detection on social media by exploring and implementing a BiLSTM network. The study shows that BiLSTM effectively detects hate speech by leveraging its ability to capture contextual dependencies and analyze sequential text input. A comprehensive evaluation of the BiLSTM model using measures such as precision, confidence, recall, and F-score confirms its effectiveness in detecting hate speech while reducing both false positives and false negatives. Furthermore, a comparison of BiLSTM with alternative machine learning algorithms illustrates its superior performance, especially in the early stages of training, indicating its fast adaptability and robustness. The results of this study confirm that BiLSTM is an effective tool in the ongoing quest to combat hate speech and promote safer online discourse. The continuous evolution of the digital world makes it possible to apply advanced NLP techniques such as BiLSTM to improve content moderation practices and create a more inclusive and respectful online environment. This study expands on the methodologies used to detect hate speech, highlighting the need for technology development to address the complex challenges of hate speech on social media.

Conclusion

In conclusion, investigating and deploying a BiLSTM network will enhance hate speech identification on social media. The paper demonstrates how BiLSTM, by utilizing its capacity to collect contextual dependencies and evaluate consecutive text input, accurately detects hate speech. The efficacy of the BiLSTM model in identifying hate speech while lowering false positives and false negatives is confirmed by a thorough study that takes into account metrics including precision, confidence, recall, and F-score. Additionally, when compared to other machine learning algorithms, BiLSTM performs better, particularly in the early training phases, demonstrating its quick adaptation and resilience. The study's findings support the notion that BiLSTM is a useful instrument in the continuous fight against hate speech and in favor of safer online conversation. Because of the way the digital world is always changing, it is now possible to use cutting-edge NLP approaches like BiLSTM to enhance content moderation procedures and foster an online community that is more welcoming and kind. In order to solve the complicated issues of hate speech on social media, technology development is necessary, as this study emphasizes by extending the approaches used to detect hate speech.

Acknowledgments

This study was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (grant No. AP23488900- Automatic detection of cyberbullying among young people in social networks using artificial intelligence)

References

[1] Govers J. et al. Down the rabbit hole: Detecting online extremism, radicalisation, and politicised hate speech //ACM Computing Surveys. – 2023. – Т. 55. – №. 14s. – С. 1-35.URL: https://doi.org/10.1145/3583067

[2] Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, Sultan D. et al. A Review of Machine Learning Techniques in Cyberbullying Detection //Computers, Materials & Continua. – 2023. – Т. 74. – №. 3. URL: DOI: 10.32604/cmc.2023.033682

[3] Ali M. et al. Social media content classification and community detection using deep learning and graph analytics //Technological Forecasting and Social Change. – 2023. – Т. 188. – С. 122252.URL: https://doi.org/10.1016/j.techfore.2022.122252

[4] Husain F., Uzuner O. A survey of offensive language detection for the Arabic language //ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). – 2021. – Т. 20. – №. 1. – С. 1-44. URL: https://doi.org/10.1145/3421504

[5] Babu N. V., Kanaga E. G. M. Sentiment analysis in social media data for depression detection using artificial intelligence: a review //SN computer science. – 2022. – Т. 3. – №. 1. – С. 74.

[6] Asghar M. Z. et al. Exploring deep neural networks for rumor detection //Journal of Ambient Intelligence and Humanized Computing. – 2021. – Т. 12. – С. 4315-4333.

[7] Ullah F. et al. IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic //Digital Communications and Networks. – 2024. – Т. 10. – №. 1. – С. 190-204.URL: https://doi.org/10.1016/j.dcan.2023.03.008

[8] Azzi S. A., Zribi C. B. O. From machine learning to deep learning for detecting abusive messages in arabic social media: survey and challenges //International Conference on Intelligent Systems Design and Applications. – Cham : Springer International Publishing, 2020. – С. 411-424.

[9] Ghosal S., Jain A. Hatecircle and unsupervised hate speech detection incorporating emotion and contextual semantics //ACM Transactions on Asian and Low-Resource Language Information Processing. – 2023. – Т. 22. – №. 4. – С. 1-28.URL: https://doi.org/10.1145/3576913

[10] Machová K., Mach M., Porezaný M. Deep learning in the detection of disinformation about COVID-19 in online space //Sensors. – 2022. – Т. 22. – №. 23. – С. 9319.URL: https://doi.org/10.3390/s22239319

[11] Singh J. P. et al. Attention-based LSTM network for rumor veracity estimation of tweets //Information Systems Frontiers. – 2022. – С. 1-16.

[12] Al-Ibrahim R. M., Ali M. Z., Najadat H. M. Detection of hateful social media content for arabic language //ACM Transactions on Asian and Low-Resource Language Information Processing. – 2023. – Т. 22. – №. 9. – С. 1-26. URL: https://doi.org/10.1145/3592792

[13] Chung J. Empirical evaluation of gated recurrent neural networks on sequence modeling //arXiv preprint arXiv:1412.3555. – 2014.