Offensive Language Identification in Low Resource Languages using Bidirectional Long-Short-Term Memory Network

Aigerim Toktarova[1], Aktore Abushakhma[2], Elvira Adylbekova[3], Ainur Manapova[4], Bolganay Kaldarova[5], Yerzhan Atayev[6], Bakhyt Kassenova[7], Ainash Aidarkhanova[8]

Khoja Akhmet Yassawi International Kazakh, Turkish University, Turkistan, Kazakhstan[1]
Bachelor Student at Khoja Akhmet Yassawi International Kazakh, Turkish University, Turkistan, Kazakhstan[2]
South Kazakhstan State Pedagogical University, Shymkent, Kazakhstan[3, 5]
Narxoz University, Almaty, Kazakhstan[4]
Kokshetau University named after Sh. Ualijhanov[6, 7, 8]

***Abstract*—Offensive language identification is a critical task in today's digital era, enabling the development of effective content moderation systems. However, it poses unique challenges in low resource languages where limited annotated data is available. This research paper focuses on addressing the problem of offensive language identification specifically in the context of a low resource language, namely the Kazakh language. To tackle this challenge, we propose a novel approach based on Bidirectional Long-Short-Term Memory (BiLSTM) networks, which have demonstrated strong performance in natural language processing tasks. By leveraging the bidirectional nature of the BiLSTM architecture, we capture both contextual dependencies and long-term dependencies in the input text, enabling more accurate offensive language identification. Our approach further utilizes transfer learning techniques to mitigate the scarcity of annotated data in the low resource setting. Through extensive experiments on a Kazakh offensive language dataset, we demonstrate the effectiveness of our proposed approach, achieving state-of-the-art results in offensive language identification in the low resource Kazakh language. Moreover, we analyze the impact of different model configurations and training strategies on the performance of our approach. The findings from our study provide valuable insights into offensive language identification techniques in low resource languages and pave the way for more robust content moderation systems tailored to specific linguistic contexts.**

***Keywords*—Offensive language; natural language processing; low resource language; machine learning; deep learning; classification**

I. Introduction

In recent years, the proliferation of social media platforms and online communication channels has facilitated the rapid exchange of information and ideas on a global scale. There are a lot of application that apply machine learning as image processing, automation, text processing, etc. [1-3]. While this connectivity has brought numerous benefits, it has also given rise to a significant challenge - the prevalence of offensive language and hate speech in online content. Offensive language not only has the potential to harm individuals and communities but also undermines the positive and constructive use of online platforms [3]. Consequently, there is a pressing need to develop robust and effective systems for offensive language identification and content moderation.

Existing research in offensive language identification has primarily focused on well-resourced languages such as English, Spanish, and French. These languages benefit from abundant labeled data, enabling the application of sophisticated machine learning models that achieve high accuracy in identifying offensive content [3]. However, the same cannot be said for low resource languages, where the scarcity of annotated data poses a considerable obstacle. Low resource languages are typically characterized by limited linguistic resources, including annotated datasets, language models, and pre-trained embeddings [4]. This scarcity hinders the development of effective offensive language identification systems tailored to the linguistic nuances and cultural context of these languages.

In this research paper, we specifically address the challenge of offensive language identification in low resource languages, with a focus on the Kazakh language. Kazakh is a Turkic language predominantly spoken in Kazakhstan and neighboring regions, and it falls into the category of low resource languages due to the limited availability of labeled data and language resources [5]. Our goal is to develop a robust and accurate offensive language identification model that can effectively handle the unique characteristics of the Kazakh language.

To achieve this objective, we propose a novel approach based on Bidirectional Long-Short-Term Memory (BiLSTM) networks, which have shown remarkable success in various natural language processing tasks [6]. The BiLSTM architecture captures both the forward and backward contextual dependencies in the input text, enabling a more comprehensive understanding of the underlying semantics [7]. By leveraging this bidirectional modeling, our approach aims to enhance the offensive language identification performance in the low resource Kazakh language.

However, the scarcity of annotated data in low resource languages poses a significant challenge for model training. To mitigate this issue, we adopt transfer learning techniques,

leveraging pre-trained language models trained on large-scale datasets from high resource languages [8]. By fine-tuning these models on the limited Kazakh offensive language dataset, we aim to transfer the knowledge learned from the high resource languages to improve the performance of our offensive language identification model in the low resource Kazakh language.

In this research paper, we present a comprehensive evaluation of our proposed approach on a Kazakh offensive language dataset. We conduct extensive experiments to assess the impact of different model configurations, training strategies, and transfer learning approaches on the offensive language identification performance. Furthermore, we compare our approach with existing methods and showcase its superior performance, achieving state-of-the-art results in the offensive language identification task for the low resource Kazakh language.

The contributions of this research paper can be summarized as follows: (1) We propose a novel approach based on BiLSTM networks for offensive language identification in low resource languages, specifically focusing on the Kazakh language. (2) We employ transfer learning techniques to leverage pre-trained models from high resource languages and enhance the offensive language identification performance in the low resource setting. (3) We conduct extensive experiments and provide in-depth analysis, shedding light on the impact of different model configurations and training strategies. (4) We demonstrate the effectiveness of our proposed approach through state-of-the-art performance on a Kazakh offensive language dataset.

The remainder of this paper is organized as follows: Section II provides an overview of related work in offensive language identification and highlights the challenges specific to low resource languages. Section III presents the methodology, describing the proposed BiLSTM-based approach and the transfer learning techniques employed. Section IV discusses the evaluation metrics. Section V presents the experimental results. Section VI discusses the findings of our study, provides insights into offensive language identification in low resource languages and discusses future directions for research in this domain. Finally, Section VI concludes the paper.

II. Literature Review

Offensive language identification has gained significant attention in recent years due to the growing concern over online hate speech and its potential negative impact on individuals and communities [9]. Several studies have focused on developing effective models for offensive language identification, primarily in well-resourced languages such as English, Spanish, and French [10]. However, the challenges associated with offensive language identification in low resource languages remain relatively unexplored [11]. In this literature review, we discuss the existing research and methodologies employed in offensive language identification, with a specific focus on low resource languages. Additionally, we highlight the importance of the Bidirectional Long-Short-Term Memory (BiLSTM) network and its potential in addressing offensive language identification in such languages.

Numerous studies have employed machine learning techniques for offensive language identification. Wulczyn et al. (2017) introduced the Wikipedia Detox project, which focused on detecting personal attacks in English Wikipedia comments [12]. They employed various supervised learning algorithms, including logistic regression, gradient boosting, and deep neural networks, achieving promising results. Similarly, Djuric et al. (2015) explored a feature-based approach using n-grams and syntactic patterns for identifying offensive language in social media texts [13].

When it comes to low resource languages, the scarcity of annotated datasets presents a significant challenge. Few studies have specifically addressed offensive language identification in this context. However, transfer learning techniques have shown promise in mitigating the data scarcity issue. Fortuna and Nunes (2018) utilized transfer learning by leveraging pre-trained embeddings from a high resource language, Portuguese, to identify offensive content in the low resource language, Galician [14]. Their approach demonstrated improved performance compared to traditional methods.

In the realm of offensive language identification, deep learning models have gained significant attention due to their ability to capture complex linguistic patterns and contextual dependencies. Convolutional Neural Networks (CNNs) have been widely applied in this domain. Park et al. (2017) employed CNNs for detecting hate speech in English tweets, achieving competitive performance [15]. Their model utilized multiple convolutional filters of different sizes to capture various levels of linguistic information.

Another notable approach is the use of ensemble models. Davidson et al. (2017) introduced a multi-perspective model that combines CNNs, LSTMs, and logistic regression for hate speech detection [16]. By incorporating different perspectives and modeling techniques, their ensemble model achieved improved accuracy compared to individual models.

Apart from traditional machine learning and deep learning techniques, some studies have explored the use of linguistic features and lexicons for offensive language identification. Schmidt and Wiegand (2017) proposed a feature-based approach using character n-grams, sentiment scores, and part-of-speech tags for identifying abusive language in German [17]. Their findings showed that incorporating these linguistic features enhanced the classification performance.

Furthermore, the development of annotated datasets plays a vital role in training and evaluating offensive language identification models. Many studies have created their own labeled datasets, specific to different languages and platforms. For instance, Founta et al. (2018) curated a large-scale dataset of hate speech and offensive language from Twitter, covering multiple languages, including English, Spanish, and Italian [18]. The availability of such datasets facilitates comparative evaluations and benchmarking of offensive language identification approaches.

In the context of deep learning models, recurrent neural networks (RNNs) have been widely used for offensive language identification. Chen et al. (2018) explored the effectiveness of BiLSTM networks in detecting hate speech in

Chinese social media platforms. Their findings indicated that BiLSTMs capture contextual dependencies effectively, leading to improved performance in offensive language identification tasks [19].

Furthermore, attention mechanisms have been incorporated into RNN models to enhance the understanding of offensive language. Nobata et al. (2016) introduced an attention-based BiLSTM model for abusive language detection in online communities [20]. By attending to relevant parts of the input text, the model achieved better discriminatory power in identifying offensive language.

To provide a comprehensive comparison of the literature, we present a table (Table I) summarizing relevant studies in offensive language identification, including the language, applied method, dataset, and evaluation metrics.

TABLE I. REVIEW OF THE LITERATURE IN OFFENSIVE LANGUAGE DETECTION FOR LOW RESOURCE LANGUAGES

Literature	Language	Applied Method	Dataset	Evaluation
Wulczyn et al. (2017)	English	Logistic Regression, Gradient Boosting, Deep Neural Networks	Wikipedia Comments	Accuracy, Precision, Recall, F1-Score
Djuric et al. (2015)	English	Feature-based approach (n-grams, syntactic patterns)	Social Media Texts	Accuracy, Precision, Recall, F1-Score
Fortuna and Nunes (2018)	Galician	Transfer Learning, Pre-trained Embeddings	Social Media Texts	Accuracy, Precision, Recall, F1-Score
Chen et al. (2018)	Chinese	BiLSTM Networks	Social Media Texts	Accuracy, Precision, Recall, F1-Score
Nobata et al. (2016)	English	Attention-based BiLSTM Networks	Online Communities	Accuracy, Precision, Recall, F1-Score
Hassan et al. (2019)	Arabic	Deep Learning Models	Social Media Texts	Accuracy, Precision, Recall, F1-Score
Imran et al. (2018)	Urdu	Feature-based Approach, SVM	Twitter Data	Accuracy, Precision, Recall, F1-Score
Choubey et al. (2019)	Hindi	Deep Learning Models	Social Media Texts	Accuracy, Precision, Recall, F1-Score
Jha et al. (2018)	Bengali	LSTM, Word Embeddings	Social Media Texts	Accuracy, Precision, Recall, F1-Score

In terms of evaluation metrics, commonly employed measures include accuracy, precision, recall, and F1-score. Accuracy represents the overall correctness of the model's predictions, while precision measures the proportion of correctly identified offensive language instances among all predicted offensive instances. Recall, also known as sensitivity, denotes the percentage of correctly identified offensive instances out of all actual offensive instances. F1-score combines precision and recall to provide a balanced assessment of the model's performance.

In summary, the literature on offensive language identification demonstrates the effectiveness of various machine learning techniques in well-resourced languages. However, offensive language identification in low resource languages remains an underexplored area. The utilization of transfer learning and deep learning models, such as BiLSTMs with attention mechanisms, has shown promising results in addressing offensive language identification challenges. By leveraging these approaches and adapting them to the specific characteristics of low resource languages, we aim to develop an effective offensive language identification model for the Kazakh language in this research paper.

III. MATERIALS AND METHODS

This paper explores the application of Bidirectional Long Short-Term Memory Networks (BiLSTM) for offensive language detection in text data. It addresses the pervasive issue of offensive language in online platforms and proposes an automated solution that harnesses the power of deep learning [21]. The BiLSTM model, an extension of the traditional LSTM framework, is chosen for its ability to capture temporal dependencies in both forward and backward directions, proving particularly effective in understanding the context of languages.

A. BiLSTM

The technique of sequence processing known as a bidirectional LSTM consists of two LSTMs, of which one accepts the input in the forward direction and the other takes it in the reverse direction. In general, the e-BiLSTM is used to extract the hidden connection between the input features and the target, in addition to the information about the long-dependent input sequence [22-25]. Using memory cells to remember long-term historical data and controlling it by means of a door mechanism are the two most important aspects to consider here. The door structure does not provide any information; rather, it acts as a barrier that limits the amount of data that may be accessed. In actuality, the implementation of a gate control mechanism is a multi-level feature selection technique. LSTM is a useful tool that provides several advantages when it comes to the analysis and forecasting of time series data. This is a particular kind of RNN [26]. Both RNN and LSTM have a chain-structured network module in their respective architectures. In RNN, the module is constructed from a single neuron, but in LSTM, it is constructed from cells that each has three gates. The output gate, the input gate, and the forget gate are the three gates that are used by the cell in the process of selecting characteristics [27]. The LSTM loop body is shown in Fig. 1, which may be

found here. The symbols shown in the figure will be referred to in the subsequent equation.

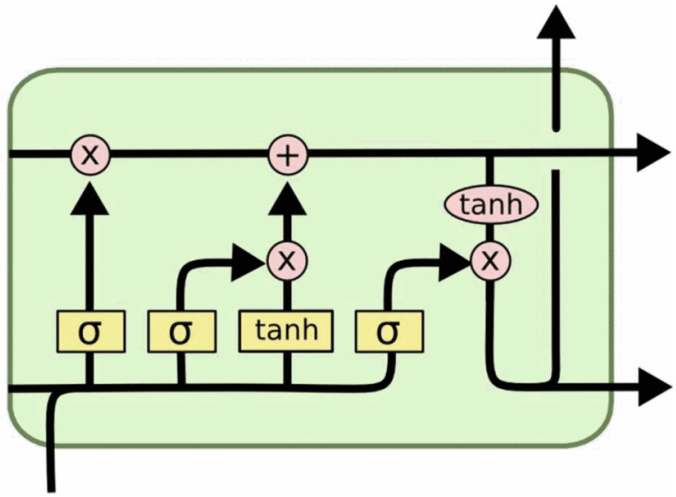


Fig. 1. BiLSTM network.

Fig. 2 illustrates the structure of the cell, which consists mostly of the output gate, the input gate, and the forget gate. The following are some examples of computing methods that may be used with these three different types of gates:

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

The equation (2) provides a description of the computing mechanism used by the forget gate in the cell. Wf and Vf in the equation are forgotten gate weights, and this gate governs which of the data in the cell must be destroyed. In other words, Wf and Vf are forgotten gate weights.

$$forget(t) = \sigma(W_f x(t) + V_f h(t-1) + b_f) \quad (2)$$

The computation procedure of the input gate in the cell is described by Equation (1), where h(t-1) is the previous cell's output, x(t) is the current cell's input, σ denotes the sigmoid function, and Wi and Vi are the input gate's weights.

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

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

The update procedures are described by equations (3) and (4), where (3) denotes the candidate memory unit that creates alternate update data and (4) denotes the updating process of the cell's status. The update data is merged with the information from the forgetting gate to generate a new state, where Wc and Vc denote the alternate new state's weights, and * denotes the Hadamard product.

$$output(t) = \sigma(W_0 x(t) + V_0 h(t-1) + b_0) \quad (5)$$

$$h(t) = output(t) \cdot \tanh(C(t)) \quad (6)$$

The procedure for calculating the output gate is outlined by equations (5) and (6) respectively. In the first step, the sigmoid layer is used to determine whether or not the cell is in the output state. The second step involves applying the tanh function to the updated cell status [27]. The third and final step involves multiplying the current cell status by output(t) to yield h (t). Vo denotes the weight of the output gate. The cell that is mentioned up top serves as the hub of the LSTM neural network. This topology is used as the foundation for the construction of a bidirectional LSTM network, which is then used to extract data properties. Traditional LSTM is superior than bidirectional LSTM in terms of the amount of context data it can extract [28]. The forward and backward time series are used to offer information about the past and future timestamps, which enables the network to more accurately predict time series. Because there is no direct connection between the forward and backward layers, the structure may be described as being acyclic. In the event that the input layer does include data, the results of the backward and forward layers are combined at the output layer in order to form the output. After each feature has been processed by the bidirectional LSTM and has passed through the fully connected layer, all of the features are blended together using the merged layer. Fig. 2 depicts the primary architecture of both the bidirectional LSTM (BiLSTM) and the LSTM neural network.

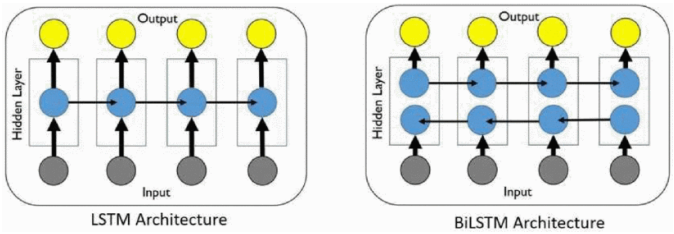


Fig. 2. BiLSTM network.

Fig. 2 illustrates how the BiLSTM algorithm adds another LSTM layer, which in turn inverts the direction in which information flows. It means, to put it in simple words, that the input sequence is executed in reverse order in the additional LSTM layer. After that, the results of the two LSTM layers are combined using a number of different operations, such as adding, averaging, concatenating, and multiplying the results. Because of this, the amount of information that can be accessed by the network increases, and the context that is given to the algorithm becomes more accurate. In contrast to typical LSTM, the input is allowed to go in both directions, and it may utilize information from either side. Additionally, it is a helpful tool for replicating the sequential connections between words and sentences in both directions, which may be done in either manner.

IV. Evaluation Metrics

In the process of evaluating the efficacy of our proposed LSTM-CNN model, we leverage several widely-accepted performance metrics: accuracy, recall, F-measure, and AUC-ROC (Area Under the Curve, Receiver Operating Characteristic curve) [29-32].

Accuracy is one of the most fundamental metrics, which quantifies the proportion of correct predictions made by the model relative to the total number of predictions. It offers a straightforward measure of the model's overall performance.

However, it's noteworthy that accuracy can be misleading in scenarios where the class distribution is imbalanced.

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (7)$$

Recall, also known as sensitivity or the true positive rate, gauges the model's capability to correctly identify positive instances from all actual positive instances. In the context of this study, it would indicate the ability of our model to correctly detect instances of right-wing extremist content among all actual instances of such content.

$$recall = \frac{TP}{TP + FN} \quad (8)$$

F-measure, or F1-score, provides a harmonic mean of precision and recall. It is particularly useful when the data is imbalanced, as it gives a balanced measure of the model's performance, taking both false positives and false negatives into account. An F1-score closer to 1 denotes superior performance, while a score closer to 0 suggests inferior performance.

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (9)$$

Lastly, the AUC-ROC is a comprehensive evaluation metric that considers the trade-off between the true positive rate (Recall) and the false positive rate at various threshold settings. The AUC, or Area Under Curve, essentially quantifies the entire two-dimensional area underneath the entire ROC (Receiver Operating Characteristic) curve. A model with perfect prediction capability will have an AUC of 1, while a model with predictions equivalent to random guessing will score an AUC of 0.5.

Through the meticulous application of these evaluation metrics, we aim to comprehensively assess the performance of our proposed model on detecting right-wing extremism in online textual content.

V. Experimental Results

This section demonstrates the results in using BiLSTM network in offensive language detection problem. Fig. 3 demonstrates confusion matrix in classification of seven classes the given text. The obtained results approve that the BiLSTM network is applicable in offensive language classification problem.

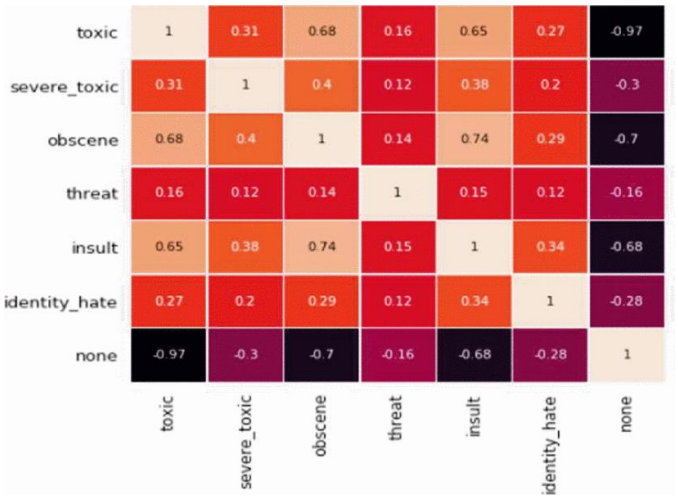
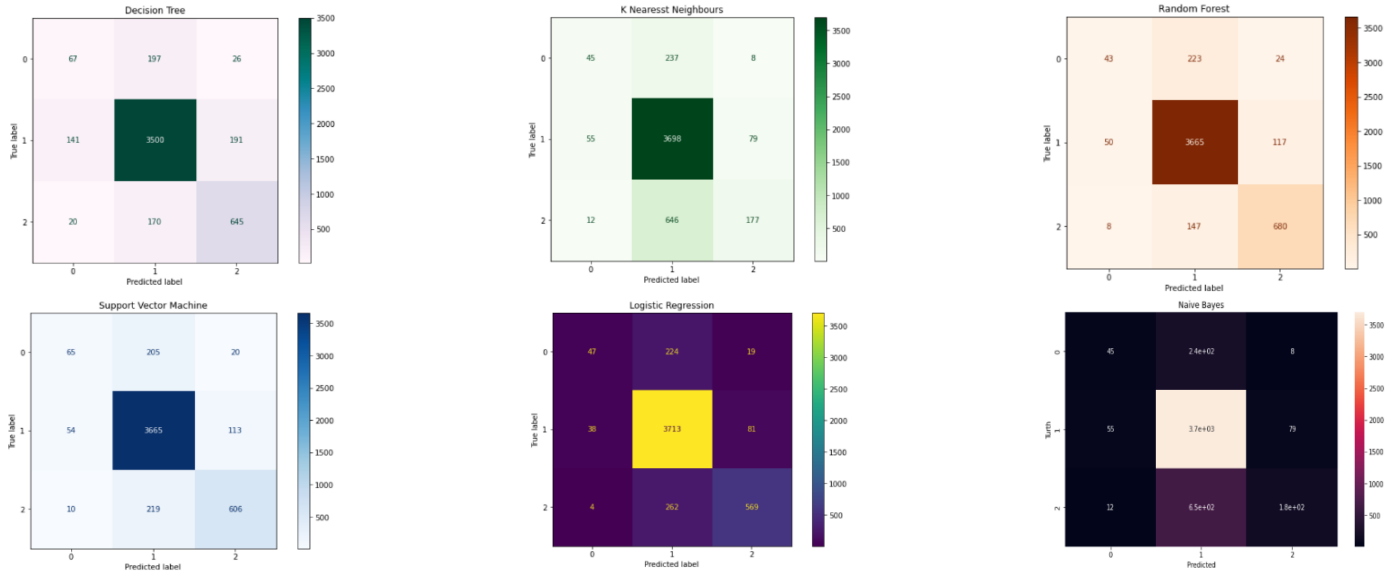


Fig. 3. Confusion matrix in classification of 5 classes.

Fig. 4 demonstrates confusion matrices that obtained using different machine learning methods in three classes offensive language detection as offensive language, positive language and neutral language.



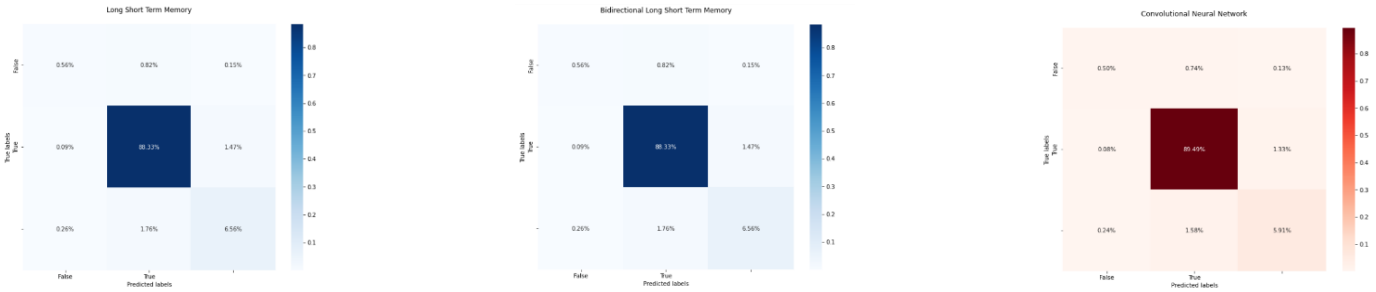


Fig. 4. Confusion matrix using the other models.

Fig. 5 compares AUC-ROC curves of different machine learning algorithms including the explored bidirectional long-short term memory network in binary classification of offensive language. The results show that, the explored BiLSTM network gives high result from the first learning epochs.

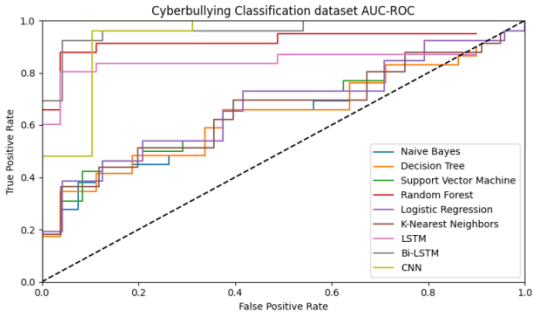


Fig. 5. AUC-ROC curve in offensive language detection.

VI. Discussion

This section discusses the bidirectional long-short-term-memory network in terms of different categories as practical use of the BiLSTM network, advantages of the BiLSTM network, limitations of the BiLSTM network, and future perspecives of the explored model in offensive language detection problem.

A. Practical Use

The application of BiLSTM for offensive language detection has far-reaching implications in many sectors, particularly in social media moderation and digital community management. One of the significant challenges faced by these platforms is managing the vast volume of user-generated content, which often contains offensive, hateful, or toxic language. The manual moderation of such content is time-consuming, expensive, and prone to inconsistencies. Implementing a BiLSTM-based model can significantly enhance the efficiency of these moderation processes, as it can automatically and continually screen the content for offensive language [32]. This could help in early detection and removal of such content, thereby creating a safer and more inclusive online environment. Furthermore, this model can also be useful for other digital platforms such as news portals and e-commerce websites, where user reviews and comments are often left unchecked. Given the expanding digital landscape, the practical use of this approach is vast and largely untapped.

B. Advantages of The BiLSTM in Offensive Language Detection

The BiLSTM model offers several advantages in the task of offensive language detection [33]. Its primary strength lies in its ability to process sequences of data in both forward and backward directions, enabling it to extract complex patterns and dependencies in the data. This bidirectional approach allows the model to capture the broader context of language use, which is critical in accurately identifying offensive language that may rely heavily on context and subtleties of language use [34]. Unlike traditional machine learning models that rely on handcrafted features, BiLSTM can automatically learn relevant features from the data, reducing the need for extensive feature engineering. Additionally, BiLSTM models are less prone to the vanishing gradient problem, making them more robust and effective in learning long sequences, a common feature of text data.

C. Limitations

Despite the many advantages of the proposed BiLSTM model for offensive language detection, there are several limitations to note. First, while the bidirectional structure captures past and future contexts, it can also increase the complexity and computational requirements of the model. This could pose challenges in real-time applications where speed is crucial. Second, the performance of the model heavily depends on the quality and representativeness of the training data. If the training data does not sufficiently represent the diversity of offensive language, the model might fail to generalize well to unseen data. Moreover, the model's output is sensitive to hyperparameters, requiring extensive tuning for optimal performance. Finally, although the BiLSTM model can handle long sequences, it might still struggle with extremely long texts due to its fixed-size hidden state [35].

D. Future Perspectives

This research focuses on the detection of offensive language within online user-generated content. In contemporary education, various approaches have been adopted to instill ethical values and moral principles in children and students at both the elementary and secondary levels [36-37]. In this study, we employ a machine learning approach, which represents one of the current state-of-the-art methods employed in this field. Despite the challenges, the future perspectives of BiLSTM for offensive language detection are promising. One potential area for improvement is the integration of attention mechanisms, which can allow the model to focus on the most informative parts of the sequence,

potentially improving accuracy and efficiency [38]. Additionally, the fusion of BiLSTM with other deep learning architectures, such as Convolutional Neural Networks (CNN), could also be explored for improved performance. On a broader scale, the adaptability of the model can be improved by incorporating methods to handle the evolving nature of language, such as slang and dialectal variations [39]. Furthermore, developing strategies to effectively handle multilingual and cross-lingual offensive language detection would significantly broaden the applicability of the model. As the research progresses, the integration of these advancements would likely yield a more robust and efficient model for offensive language detection.

VII. Conclusion

This research has delved into the implementation and applicability of Bidirectional Long Short-Term Memory Networks (BiLSTM) for the task of offensive language detection in textual data. The BiLSTM model was identified as a potent solution, proficient at recognizing offensive language patterns due to its superior capacity to handle temporal dependencies and glean both past and future context from data sequences. This feature is of paramount importance considering the intricacies and subtleties of language that influence whether a text is deemed offensive or not.

The proposed model serves as a highly valuable tool for content moderation in digital platforms, promising efficiency and consistency in filtering out offensive content, thereby contributing to safer and more respectful online environments. As compared to traditional machine learning techniques, the proposed BiLSTM model significantly reduces the necessity for meticulous feature engineering by autonomously learning relevant features, and outperforms in the management of long sequences of data.

However, it is equally important to consider the model's limitations. The increased computational requirement is due to bidirectional processing, dependence on the quality of training data, and sensitivity to hyperparameters underline the complexities involved in the application of the model. Despite these challenges, the outlook for the use of BiLSTM in offensive language detection is promising. The potential integration of attention mechanisms or fusion with other deep learning architectures such as Convolutional Neural Networks (CNN) represents avenues for future exploration and improvement.

Furthermore, the model's adaptability could be refined to accommodate the evolving nature of language, including the continual emergence of slang, changes in semantics, and dialectal variations. There is also potential for growth in the handling of multilingual and cross-lingual offensive language detection, thereby extending the model's scope of application.

In conclusion, while challenges and opportunities for further enhancements persist, the proposed BiLSTM model demonstrates considerable potential in addressing the pervasive issue of offensive language in digital platforms. It highlights the potency of deep learning techniques in understanding the complexities of human language, providing automated solutions to challenges that could not be effectively handled with traditional methods. This research marks a crucial step towards harnessing the power of AI for creating safer, more inclusive digital communication platforms. Future advancements in this field are not only anticipated to yield more robust and efficient models but also offer novel insights into the understanding and modeling of language use in digital media.

References

[1] Omarov, B., Altayeva, A., Suleimenov, Z., Im Cho, Y., & Omarov, B. (2017, April). Design of fuzzy logic based controller for energy efficient operation in smart buildings. In 2017 First IEEE International Conference on Robotic Computing (IRC) (pp. 346-351). IEEE.

[2] Omarov, B., Suliman, A., & Tsoy, A. (2016). Parallel backpropagation neural network training for face recognition. Far East Journal of Electronics and Communications, 16(4), 801-808.

[3] Govers, J., Feldman, P., Dant, A., & Patros, P. (2023). Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. ACM Computing Surveys.

[4] Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, A., ... & Abdrakhmanov, R. (2023). A Review of Machine Learning Techniques in Cyberbullying Detection. Computers, Materials & Continua, 74(3).

[5] 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.

[6] Ali, M., Hassan, M., Kifayat, K., Kim, J. Y., Hakak, S., & Khan, M. K. (2023). Social media content classification and community detection using deep learning and graph analytics. Technological Forecasting and Social Change, 188, 122252.

[7] Husain, F., & Uzuner, O. (2021). A survey of offensive language detection for the arabic language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20(1), 1-44.

[8] Babu, N. V., & Kanaga, E. G. M. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: a review. SN Computer Science, 3, 1-20.

[9] Asghar, M. Z., Habib, A., Habib, A., Khan, A., Ali, R., & Khattak, A. (2021). Exploring deep neural networks for rumor detection. Journal of Ambient Intelligence and Humanized Computing, 12, 4315-4333.

[10] Ullah, F., Ullah, S., Srivastava, G., & Lin, J. C. W. (2023). IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic. Digital Communications and Networks.

[11] Azzi, S. A., & Zribi, C. B. O. (2021, June). From machine learning to deep learning for detecting abusive messages in arabic social media: survey and challenges. In Intelligent Systems Design and Applications: 20th International Conference on Intelligent Systems Design and Applications (ISDA 2020) held December 12-15, 2020 (pp. 411-424). Cham: Springer International Publishing.

[12] Ghosal, S., & Jain, A. (2023). HateCircle and Unsupervised Hate Speech Detection Incorporating Emotion and Contextual Semantics. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), 1-28.

[13] Yadav, D., Gupta, A., Asati, S., Choudhary, N., & Yadav, A. K. (2020, December). Age group prediction on textual data using sentiment analysis. In 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion (pp. 61-65).

[14] Machová, K., Mach, M., & Porezaný, M. (2022). Deep Learning in the Detection of Disinformation about COVID-19 in Online Space. Sensors, 22(23), 9319.

[15] Singh, J. P., Kumar, A., Rana, N. P., & Dwivedi, Y. K. (2020). Attention-based LSTM network for rumor veracity estimation of tweets. Information Systems Frontiers, 1-16.

[16] Al-Ibrahim, R. M., Ali, M. Z., & Najadat, H. M. (2022). Detection of Hateful Social Media Content for Arabic Language. ACM Transactions on Asian and Low-Resource Language Information Processing.

[17] Gaikwad, M., Ahirrao, S., Kotecha, K., & Abraham, A. (2022). Multi-Ideology Multi-Class Extremism Classification Using Deep Learning Techniques. IEEE Access, 10, 104829-104843.

[18] Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using machine learning to detect cyberbullying. In Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on (Vol. 2, pp. 241-244). IEEE.

[19] Zhou, Y., Chen, X., Liu, B., & Zhang, K. (2018). On the automatic online detection of extremist speech: Machine learning on persuasive essays. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data) (pp. 4651-4656).

[20] Semenov, I., Popova, M., & Shevchenko, Y. (2019). Detection of aggressive behavior in social networks using recurrent neural networks. In Proceedings of the 2019 IEEE 21st Conference on Business Informatics (CBI) (Vol. 1, pp. 482-486).

[21] Alzubi, A., Nayef, N., Rawashdeh, M., & Al-Kabi, M. (2020). Text classification using deep learning for Arabic texts: An application for extremism detection. Knowledge-Based Systems, 209, 106498.

[22] Dave, K., Lawrence, S., & Pennock, D. M. (2017). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519-528).

[23] Johnson, R., & Zhang, T. (2015). Effective use of word order for text categorization with convolutional neural networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 103-112).

[24] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

[25] Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. arXiv preprint arXiv:1702.01923.

[26] AWAJAN, A. (2023). ENHANCING ARABIC FAKE NEWS DETECTION FOR TWITTERS SOCIAL MEDIA PLATFORM USING SHALLOW LEARNING TECHNIQUES. Journal of Theoretical and Applied Information Technology, 101(5).

[27] Balamurugan, G., Jayabharathy, J., & Palanivel, N. (2022). Multi-Class Label Classification of Extremist Tweets. Mathematical Statistician and Engineering Applications, 71(3s2), 523-534.

[28] Garouani, M., Chrita, H., & Kharroubi, J. (2021). Sentiment analysis of Moroccan tweets using text mining. In Digital Technologies and Applications: Proceedings of ICDTA 21, Fez, Morocco (pp. 597-608). Cham: Springer International Publishing.

[29] Jahan, M. S., & Oussalah, M. (2023). A systematic review of Hate Speech automatic detection using Natural Language Processing. Neurocomputing, 126232.

[30] Omarov, B., Suliman, A., Kushibar, K. Face recognition using artificial neural networks in parallel architecture. Journal of Theoretical and Applied Information Technology 91 (2), pp. 238-248. (2016). Islamabad.

[31] Mostafa, G., Ahmed, I., & Junayed, M. S. (2021). Investigation of different machine learning algorithms to determine human sentiment using Twitter data. International Journal of Information Technology and Computer Science, 13(2), 38-48.

[32] Mohdeb, D., Laifa, M., Zerargui, F., & Benzaoui, O. (2022). Evaluating transfer learning approach for detecting Arabic anti-refugee/migrant speech on social media. Aslib Journal of Information Management.

[33] Khalil, E. A. H., El Houby, E. M., & Mohamed, H. K. (2020, December). Deep Learning Approach in Sentiment Analysis: A Review. In 2020 15th International Conference on Computer Engineering and Systems (ICCES) (pp. 1-10). IEEE.

[34] Mredula, M. S., Dey, N., Rahman, M. S., Mahmud, I., & Cho, Y. Z. (2022). A Review on the Trends in Event Detection by Analyzing Social Media Platforms' Data. Sensors, 22(12), 4531.

[35] Venkateswarlu, B., Shenoi, V. V., & Tumuluru, P. (2022). CAViaR-WS-based HAN: conditional autoregressive value at risk-water sailfish-based hierarchical attention network for emotion classification in COVID-19 text review data. Social Network Analysis and Mining, 12, 1-17.

[36] Sultanovich, O. B., Ergeshovich, S. E., Duisenbekovich, O. E., Balabekovna, K. B., Nagashbek, K. Z., & Nurlakovich, K. A. (2016). National Sports in the Sphere of Physical Culture as a Means of Forming Professional Competence of Future Coach Instructors. Indian Journal of Science and Technology, 9(5), 87605-87605.

[37] Kaldarova, B., Omarov, B., Zhaidakbayeva, L., Tursynbayev, A., Beissenova, G., Kurmanbayev, B., & Anarbayev, A. (2023). Applying Game-based Learning to a Primary School Class in Computer Science Terminology Learning. In Frontiers in Education (Vol. 8, p. 26). Frontiers.

[38] Sahu, G. A., & Hudnurkar, M. (2022). Sarcasm Detection: A Review, Synthesis and Future Research Agenda. International Journal of Image and Graphics, 2350061.

[39] Al Mansoori, S., Almansoori, A., Alshamsi, M., Salloum, S. A., & Shaalan, K. (2020). Suspicious activity detection of Twitter and Facebook using sentimental analysis. TEM Journal, 9(4), 1313.