Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Innovative deep learning techniques for monitoring aggressive behavior in social media posts

Huimin Han[1], Muhammad Asif[2*], Emad Mahrous Awwad[3], Nadia Sarhan[4], Yazeed Yasid Ghadi[5] and Bo Xu[6]

Abstract

The study aims to evaluate and compare the performance of various machine learning (ML) classifiers in the context of detecting cyber-trolling behaviors. With the rising prevalence of online harassment, developing effective automated tools for aggression detection in digital communications has become imperative. This research assesses the efficacy of Random Forest, Light Gradient Boosting Machine (LightGBM), Logistic Regression, Support Vector Machine (SVM), and Naive Bayes classifiers in identifying cyber troll posts within a publicly available dataset. Each ML classifier was trained and tested on a dataset curated for the detection of cyber trolls. The performance of the classifiers was gauged using confusion matrices, which provide detailed counts of true positives, true negatives, false positives, and false negatives. These metrics were then utilized to calculate the accuracy, precision, recall, and F1 scores to better understand each model's predictive capabilities. The Random Forest classifier outperformed other models, exhibiting the highest accuracy and balanced precision-recall trade-off, as indicated by the highest true positive and true negative rates, alongside the lowest false positive and false negative rates. LightGBM, while effective, showed a tendency towards higher false predictions. Logistic Regression, SVM, and Naive Bayes displayed identical confusion matrix results, an anomaly suggesting potential data handling or model application issues that warrant further investigation. The findings underscore the effectiveness of ensemble methods, with Random Forest leading in the cyber troll detection task. The study highlights the importance of selecting appropriate ML algorithms for text classification tasks in social media contexts and emphasizes the need for further scrutiny into the anomaly observed among the Logistic Regression, SVM, and Naive Bayes results. Future work will focus on exploring the reasons behind this occurrence and the potential of deep learning techniques in enhancing detection performance.

Keywords Cyber troll detection, Machine learning, Random forest, LightGBM, Logistic regression, SVM, Naive bayes, Text classification, Online harassment

*Correspondence:
Muhammad Asif
asifanu@hotmail.com
Full list of author information is available at the end of the article

Springer Open

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

The digital era has ushered in an era of unprecedented connectivity, transforming social media into a pivotal platform for global communication and public discourse. This virtual interconnectedness, while facilitating a plethora of meaningful interactions, has also given rise to a pernicious phenomenon: cyberbullying [1]. Cyberbullying, defined as the use of digital platforms to intimidate, belittle, or harass individuals, poses a unique challenge in the realm of online safety due to the anonymity afforded by these platforms. Its implications are far-reaching, often resulting in severe psychological and emotional distress, surpassing the impact of traditional, physical bullying in its potential for harm [2]. The spectrum of cyberbullying encompasses various manifestations, including but not limited to, racism, sexism, and cyber aggression. Cyber aggression, in particular, denotes behaviors that are hostile or hateful, often motivated by discriminatory beliefs based on race, nationality, religion, gender, and other such factors [3, 4]. These digital acts of aggression are not bound by age or demographic, making them a universal concern [5]. With the voluminous flow of content on social media platforms - millions of Facebook posts and tweets generated every minute - the task of monitoring and mitigating offensive content becomes a Herculean endeavor [6]. Notably, a significant portion of these posts contains elements of offensive language or sentiment, necessitating robust mechanisms for detection and intervention [7, 8]. Conventional approaches to tackling this issue have primarily relied on machine learning models, utilizing techniques like support vector machines (SVM), logistic regression (LR), and naïve Bayes (NB) for text classification. However, these methods, focusing largely on textual features through mechanisms like term frequency-inverse document frequency (TF-IDF) and Word2Vec, often fall short of capturing the nuanced emotional context of the communications.

The emergence of innovative deep-learning techniques for monitoring aggressive behavior in social media posts represents a significant advancement in the field of digital communication and online safety. The significance of this development lies in its potential to address a critical and growing concern in the virtual landscape: the prevalence of cyber aggression and its detrimental impact on individuals and communities. As social media platforms have become integral to daily communication, they have also unfortunately become venues for harmful behaviors like harassment, bullying, and the spread of hateful rhetoric. Traditional methods for identifying and mitigating such behavior often struggle to keep pace with the sheer volume and complexity of content generated on these platforms.

Deep learning techniques, with their ability to learn and adapt from vast amounts of unstructured data, offer a promising solution [9, 10]. By employing so0isticated algorithms and neural network architectures, these techniques can effectively analyze the nuances of language, context, and sentiment present in social media posts [11, 12]. This capability enables more accurate and comprehensive identification of aggressive behavior, going beyond mere keyword recognition to understand the subtleties of human communication, such as sarcasm, irony, and indirect speech [13]. Furthermore, the application of deep learning in this context is significant for its proactive approach to online safety. It not only aids in the immediate detection and removal of harmful content but also contributes to the larger goal of fostering healthier online environments. This can have far-reaching implications, from supporting individual mental health and well-being to promoting more respectful and constructive digital discourse [14]. By advancing these techniques, researchers and practitioners are taking critical steps toward mitigating the negative impacts of the digital age, thereby enhancing the overall quality and safety of online communication. To address this gap, our research introduces an innovative deep learning-based framework for the detection of cyber aggression. Our approach leverages a combination of novel emotional features extracted from textual data, alongside conventional Word2Vec features, to enhance the accuracy of aggression detection. The proposed deep neural network (DNN) model, characterized by its optimized architecture with a minimal number of layers, sets a new standard in both efficiency and effectiveness.

This paper delineates the following contributions:

- Demonstrated the superior performance of the Random Forest algorithm over other conventional machine learning classifiers (LightGBM, Logistic Regression, SVM, and Naive Bayes) in the context of cyber troll detection, providing evidence for its robustness in handling both specificity and sensitivity within the dataset.
- Revealed a unique outcome where Logistic Regression, SVM, and Naive Bayes classifiers yielded identical confusion matrices, prompting critical discussions on model validation and highlighting the necessity for meticulous experimental setup in machine learning workflows.
- Contributed to the field of online behavior analysis by quantitatively comparing the efficacy of different machine-learning approaches, offering insights that can guide the development of more effective automated moderation tools to combat cyber trolling and enhance digital communication safety.

Following this introduction, the paper is structured as follows: Sect. 2 offers an in-depth review of existing literature on aggression detection. Section 3 elucidates the methodology and functionality of the proposed DNN algorithm. Section 4 presents the empirical findings derived from our model. Finally, Sect. 5 provides a thorough discussion of these results, alongside considerations for future research directions.

Literature review

The rapid expansion of the social web has catalyzed significant advancements in the field of Natural Language Processing (NLP), particularly in the context of analyzing and interpreting the diverse array of communications that take place on social media platforms. These platforms, including Twitter, Facebook, and various weblogs, serve as melting pots of global interaction, bringing together individuals of different languages, races, and cultural backgrounds [15]. This diversity, while enriching, also presents unique challenges, particularly in the form of cyberbullying, online aggression, and hate speech, compounded by the intricacies and complexities inherent in processing various foreign languages [16]. Researchers have used a range of terminologies to categorize and study these negative behaviors [17]. Terms such as cyberbullying, offensive language, hate speech, racism, and profanity have been extensively explored in literature. Studies have varied in their focus, with some examining the psychological profiles of cyber-aggressors versus non-aggressors, while others have utilized text, network, and user-based features for detecting aggression in social media datasets. Notably, patterns have emerged, such as bullying victims tending to write fewer posts and participate less in discussions, in contrast to aggressors who are often more active and propagate negativity online [18, 19].

The primary focus of computational linguistics has traditionally been on resource-rich languages like English, leaving resource-poor languages somewhat underexplored due to a lack of datasets and tools. Nevertheless, there have been significant efforts to detect offensive language in various languages using machine learning algorithms. These studies have applied techniques ranging from bag-of-words and basic classifiers like multinomial-naïve Bayes and logistic regression to more advanced deep learning methods. The exploration has not been limited to English, with studies extending to languages like Hindi, Marathi, Arabic, Indonesian, German, and Portuguese. In the realm of English language datasets, researchers have made notable strides in identifying cyberbullying and other forms of online aggression. Experiments have been conducted using a variety of features, including syntactic and semantic analysis, emoji usage, and sentiment lexicons. These studies have also delved into the complexities of detecting sarcasm and irony, which are particularly challenging due to their subtlety and context-dependent nature. The advent of deep learning has brought new dimensions to NLP research, proving to be more efficient in certain aspects than traditional machine learning techniques. Deep learning's strength lies in its ability to process and learn from large sets of unstructured data, making it particularly suitable for analyzing the vast and varied content found on social media. Applications have ranged from distinguishing between hate speech and profanity to performing high-level classification of text data. Techniques like convolutional neural networks (CNN), long short-term memory (LSTM) networks, bidirectional LSTM (BiLSTM), gated recurrent units (GRU), and recurrent neural networks (RNN) have been employed to great effect [20–24].

The detection of abusive behavior on online social networks has emerged as a critical area of study due to the escalating prevalence of various forms of online abuse, including offensive language, hate speech, cyberbullying, aggression, and sexual exploitation. Research efforts have been diverse, with some focusing on the identification of potential offenders in online communities, such as YouTube comment Sect. [11], while others target the detection of hate speech, with a particular emphasis on identifying racist and sexist content [25]. A notable advancement in this domain is the proposal of methodologies that incorporate user profiles, content, and network dynamics to delineate aggressive behavior on platforms like Twitter [5, 26, 27]. Machine Learning (ML)-based approaches remain at the forefront of combating online abuse. Traditional ML classifiers, including logistic regression [8, 9, 12, 27], support vector machines [28], and ensemble classifiers [29], have been extensively deployed. For example, a study on Yahoo Finance and News data applied ML methods to discern hate speech [12], while another research used an ensemble of probabilistic, rule-based, and spatial classifiers to investigate the propagation of online hate speech on Twitter [29].

In pursuit of enhanced detection efficiency, deep learning architectures have been increasingly adopted. A spectrum of deep learning models, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), and FastText [30], have undergone evaluation for their efficacy in this domain. Furthermore, a hybrid of CNN and Gated Recurrent Unit (GRU) networks, augmented with word embeddings, has been employed for hate speech detection on Twitter [25]. The use of CNNs for the same purpose has also been reported [31]. Interestingly, a comparative study indicated that

traditional machine learning methods outperformed deep neural networks, specifically Recurrent Neural Networks (RNNs), in detecting abusive and aggressive behaviors [5].

Historically, research has concentrated on "batch mode" detection of abusive behaviors, optimizing ML classifiers to identify various types of abuse within a dataset. While some methods have achieved high accuracy, they often incur significant computational costs during the training and testing phases. However, given the dynamic nature of online content, there is an imperative need for systems capable of ongoing monitoring to detect abusive behavior promptly.

To address this, an "incremental computation" approach has been proposed, which utilizes data from preceding stages to enhance the efficiency of feature extraction and classification processes [24]. Additionally, an online framework designed for real-time cyberbullying detection on Instagram employs an online feature selection technique to maintain scalability by optimizing the feature set used for classification [14]. These methods, however, concentrate on media session-level analysis rather than individual content pieces, contrasting with approaches that target aggression detection on a per-item basis, such as individual tweets. In summary, the literature reflects a growing recognition of the complexity and multifaceted nature of online aggression and the need for sophisticated, nuanced approaches to detect and mitigate it. The evolution from basic machine learning to more advanced deep learning techniques underscores the ongoing efforts to effectively analyze and understand the rich tapestry of human communication in the digital sphere.

Methodology

Figure 1 shows the proposed model used in this study.

Models

Logistic regression

Application: Logistic Regression is a widely used classification algorithm. In the context of aggression detection, it can be applied to predict whether a social media post is cyber-aggressive or non-cyber-aggressive based on features extracted from the text. The chosen settings, including L2 regularization and lbfgs solver, help mitigate overfitting and enhance model stability.

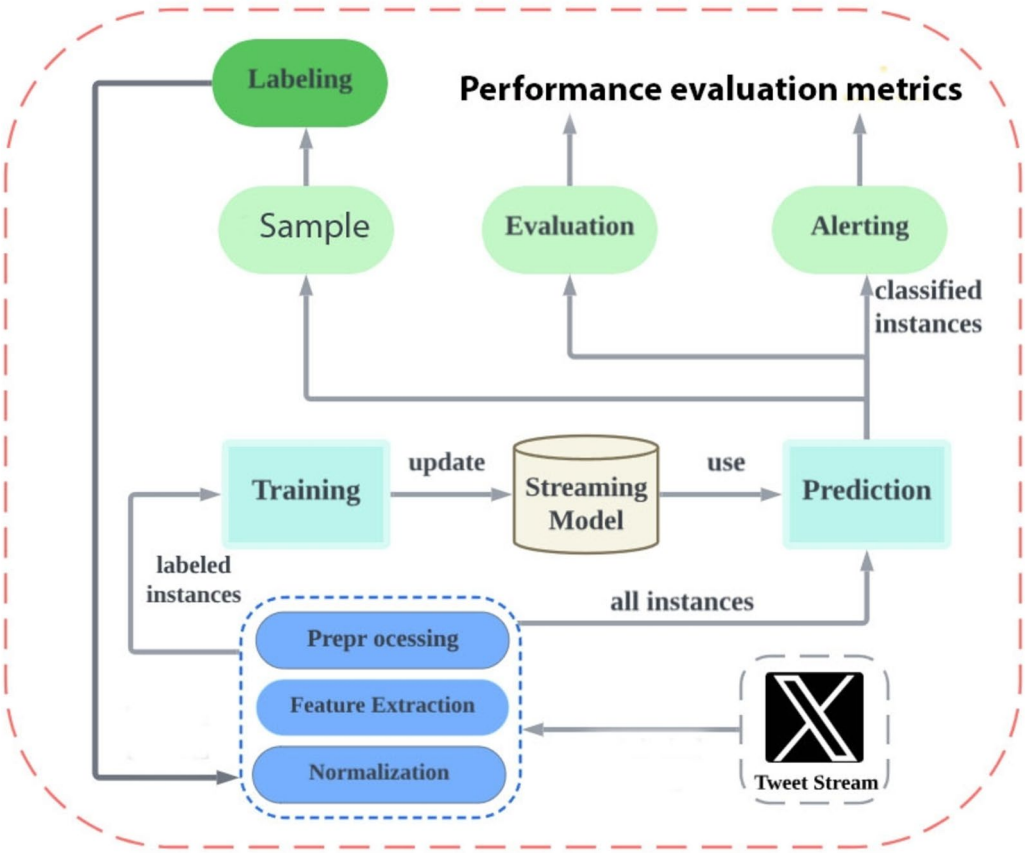

Fig. 1 Proposed model of this study

Support Vector Machine (SVM)

Application: SVM is effective for binary classification tasks. In aggression detection, SVM with the RBF kernel can capture complex relationships between features. The chosen settings, such as the RBF kernel and probability estimation, enable the model to handle non-linear decision boundaries and provide probability scores, aiding in the confidence estimation of predictions.

Naive bayes

Application: Naive Bayes is a probabilistic algorithm suitable for text classification. In aggression detection, it can model the probability of a post being cyber-aggressive or non-cyber-aggressive based on the occurrence of words. The chosen settings, including additive smoothing (alpha) and fit_prior, contribute to a robust model, particularly in dealing with sparse data.

Random forest

Application: Random Forest is an ensemble learning method known for its robustness and ability to handle complex relationships. In aggression detection, it can be used to aggregate predictions from multiple decision trees. The settings, such as the number of estimators and minimum samples for splitting, influence the model's capacity to generalize and capture patterns effectively.

LightGBM

Application: LightGBM is a gradient-boosting framework that excels in handling large datasets. In aggression detection, it can efficiently capture complex dependencies in the data. The specified settings, including binary classification as the objective and parameters controlling tree structure (num_leaves), learning rate, and feature/bagging fractions, contribute to model efficiency and accuracy.

Dataset

The dataset used in this research is the Cyber-Troll dataset, which is publicly available on Kaggle (https://www.kaggle.com/datasets/dataturks/dataset-for-detection-of-cybertrolls) and was accessed on February 9, 2022. This dataset was curated by Data-Turk for aggression detection, specifically focusing on cyber-aggressive behavior in English-language tweets.

The dataset consists of a total of 20,001 tweets, each labeled into one of two classes: cyber-aggressive (CA) and non-cyber-aggressive (NCA). The labels were assigned by the Data-Turk society based on the content of the tweets. Cyber-aggressive tweets are those that contain messages intended to insult or harm someone online, while non-cyber-aggressive tweets are those that do not carry any negative meaning and are not directed toward causing harm to others.

The distribution of the dataset is as follows:

Non-cyber-aggressive (NCA) tweets: 12,179 tweets.

Cyber-aggressive (CA) tweets: 7,822 tweets.

This distribution indicates that approximately 39% of the dataset consists of cyber-aggressive tweets, while the remaining 61% comprises non-cyber-aggressive tweets. The dataset serves as a valuable resource for training and evaluating models aimed at the detection of cyber-aggressive behavior in social media contexts. The imbalanced nature of the dataset, with a higher proportion of non-cyber-aggressive tweets, should be taken into consideration when designing and evaluating models to ensure robust and accurate performance across both classes.

Parameter settings

Logistic regression

C (Inverse of regularization strength): 1.0.

Penalty: L2 regularization.

Solver: lbfgs (Limited-memory Broyden–Fletcher–Goldfarb–Shanno).

Max Iterations: 100.

Random State: 42 (for reproducibility).

Support Vector Machine (SVM)

C (Regularization parameter): 1.0.

Kernel: RBF (Radial Basis Function).

Gamma: Scale (kernel coefficient).

Degree: 3 (degree of the polynomial kernel function).

Probability: True (to enable probability estimates).

Random State: 42 (for reproducibility).

Naive bayes

Alpha: 1.0 (Additive smoothing parameter).

Fit Prior: True (whether to learn class prior probabilities).

Random Forest:

N Estimators: 100 (Number of trees in the forest).

Max Depth: None (Maximum depth of the tree).

Min Samples Split: 2 (Minimum number of samples required to split an internal node).

Random State: 42 (for reproducibility).

LightGBM

Objective: Binary (binary classification).

Boosting Type: gbdt (Gradient Boosting Decision Tree).

Num Leaves: 31 (maximum number of leaves in one tree).

Learning Rate: 0.05 (shrinkage rate to prevent overfitting).

Feature Fraction: 0.9 (fraction of features to be used for each boosting round).

Bagging Fraction: 0.8 (fraction of data to be randomly sampled for bagging).

Bagging Freq: 5 (frequency for bagging).

Metric: Binary Logloss (logarithmic loss for binary classification).

Random State: 42 (for reproducibility).

These parameter settings provide a specific configuration for each algorithm, influencing their behavior during the training and prediction phases. Adjusting these parameters allows fine-tuning of the models to achieve optimal performance on the given task or dataset. The use of a consistent random state (42) helps in obtaining reproducible results across different runs.

Performance evaluation

The evaluation metrics employed in this study encompass average accuracy, recall, precision, and F1-score. The computation of these metrics relies on the enumeration of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) instances. True positives (TP) signify accurately classified cyber-aggressive tweets, while false negatives (FN) represent tweets erroneously categorized as non-cyber-aggressive. True negatives (TN) denote correctly classified non-cyber-aggressive tweets, while false positives (FP) correspond to tweets inaccurately labeled as cyber-aggressive.

Accuracy, a fundamental metric, is determined by the ratio of correctly classified cyber-aggressive and non-aggressive tweets to the total dataset. It serves as a holistic indicator of overall model performance. The computation of recall, precision, and F1-score involves specific aspects of classification outcomes.

Recall, or sensitivity, quantifies the proportion of actual cyber-aggressive tweets correctly identified by the model, emphasizing the model's ability to capture all instances of cyber-aggression. Precision gauges the accuracy of the model in correctly identifying cyber-aggressive tweets among those it categorizes as such, minimizing false positives. F1-score, a harmonic mean of precision and recall, offers a balanced assessment of a model's performance by considering both false positives and false negatives.

Precision measures the number of correctly identified cyber aggression tweets among all tweets labeled as cyber-aggressive.

The recall is the number of aggressive tweets among all of the tweets in the dataset.

F1-score is a measure of how well your classifier balances precision and recall.

Results

Figure 2. shows the confusion matrices for a set of machine learning classifiers, namely Random Forest, LightGBM (Light Gradient Boosting Machine), Logistic Regression, SVM (Support Vector Machine), and Naive Bayes. Confusion matrices are critical in machine learning for quantifying the performance of classification algorithms, as they provide a detailed breakdown of correct and incorrect predictions concerning actual outcomes.

The Random Forest classifier exhibits a superior predictive performance, as evidenced by the highest number of true positives (TP = 1802) and true negatives (TN = 2933), coupled with the lowest numbers of false positives (FP = 114) and false negatives (FN = 152). This suggests a robust ability to discriminate between the classes with both high sensitivity (as indicated by the high TP rate) and high specificity (as indicated by the high TN rate).

In contrast, the LightGBM classifier demonstrates a higher number of both false positives (FP = 536) and false negatives (FN = 576), indicative of a lower specificity and sensitivity respectively compared to the Random Forest classifier. The higher FP rate might suggest a tendency towards over-predicting the positive class, while the higher FN rate might indicate a conservative stance on predicting the positive class, requiring a stronger signal or evidence.

Interestingly, the confusion matrices for Logistic Regression, SVM, and Naive Bayes are identical, which may raise questions about the experimental setup or data partitioning, as it is uncommon for distinct models to yield the exact same confusion matrix on non-trivial tasks. Nevertheless, taken at face value, these classifiers have balanced false positive and false negative rates (FP = FN = 557 and 707 respectively), but they are outperformed by the Random Forest classifier in all aspects of the confusion matrix.

Figure 3 shows the comparison of performance metrics for four different machine learning models applied to a dataset for the detection of cyber trolls, as per the link you mentioned. The models evaluated are Logistic Regression, SVM (Support Vector Machine), Naive Bayes, and Random Forest.

Each model is evaluated on four different metrics:

Accuracy

This metric shows how often the model is correct when predicting whether a post is aggressive or not.

Precision

This indicates the proportion of posts that the model correctly identified as aggressive out of all the posts it labeled as aggressive.

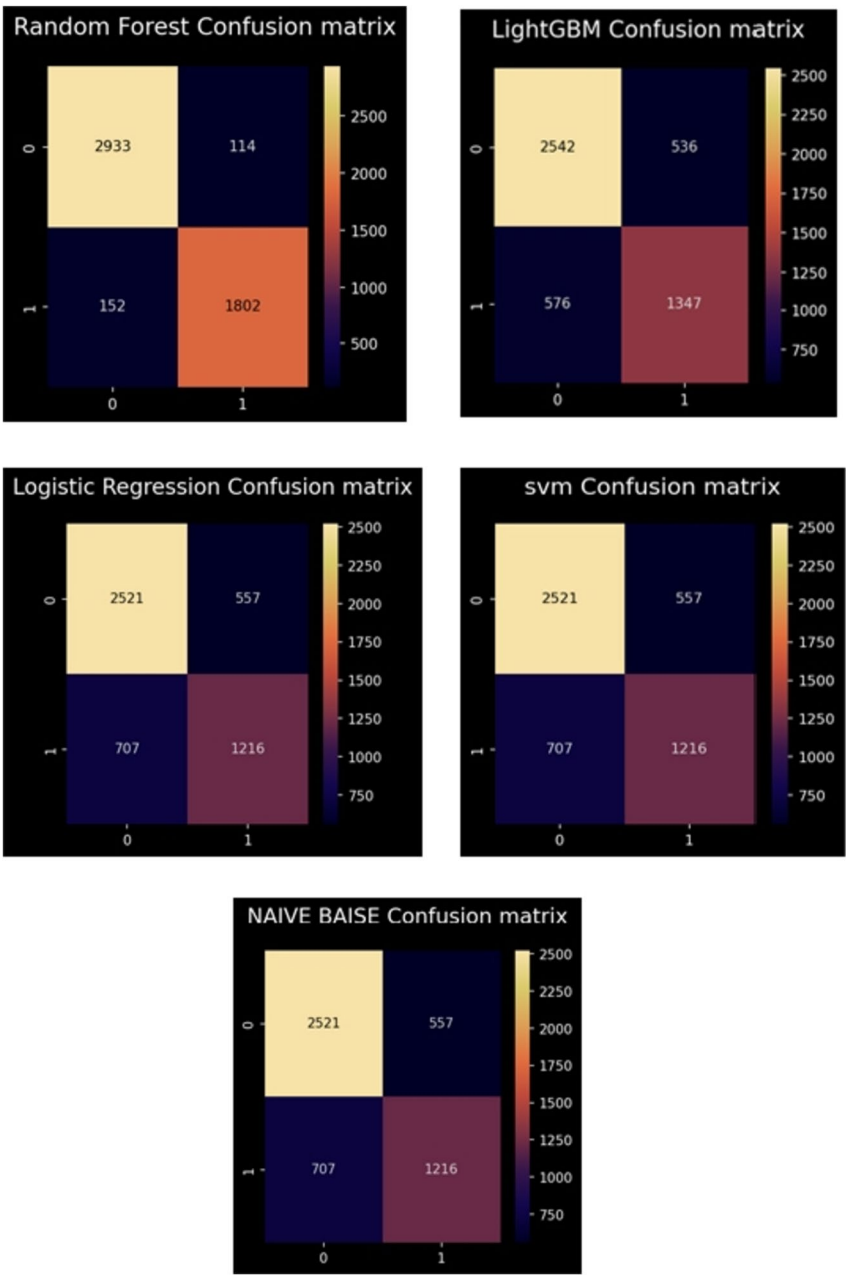

Fig. 2 Performance comparison of different models using a confusion matrix

Recall

This tells us what proportion of actual aggressive posts were correctly identified by the model.

F1 score

This is the harmonic mean of precision and recall, providing a single score that balances the two other metrics.

From the graph, we can see the performance of each model on these metrics without referring to the colors:

The Random Forest model has the highest bars across all four metrics, suggesting it has the best overall performance for detecting aggression in posts in the dataset.

The SVM model appears to perform second best, with bars slightly lower than Random Forest in all metrics.

The Logistic Regression model has lower metrics in comparison to SVM and Random Forest, particularly

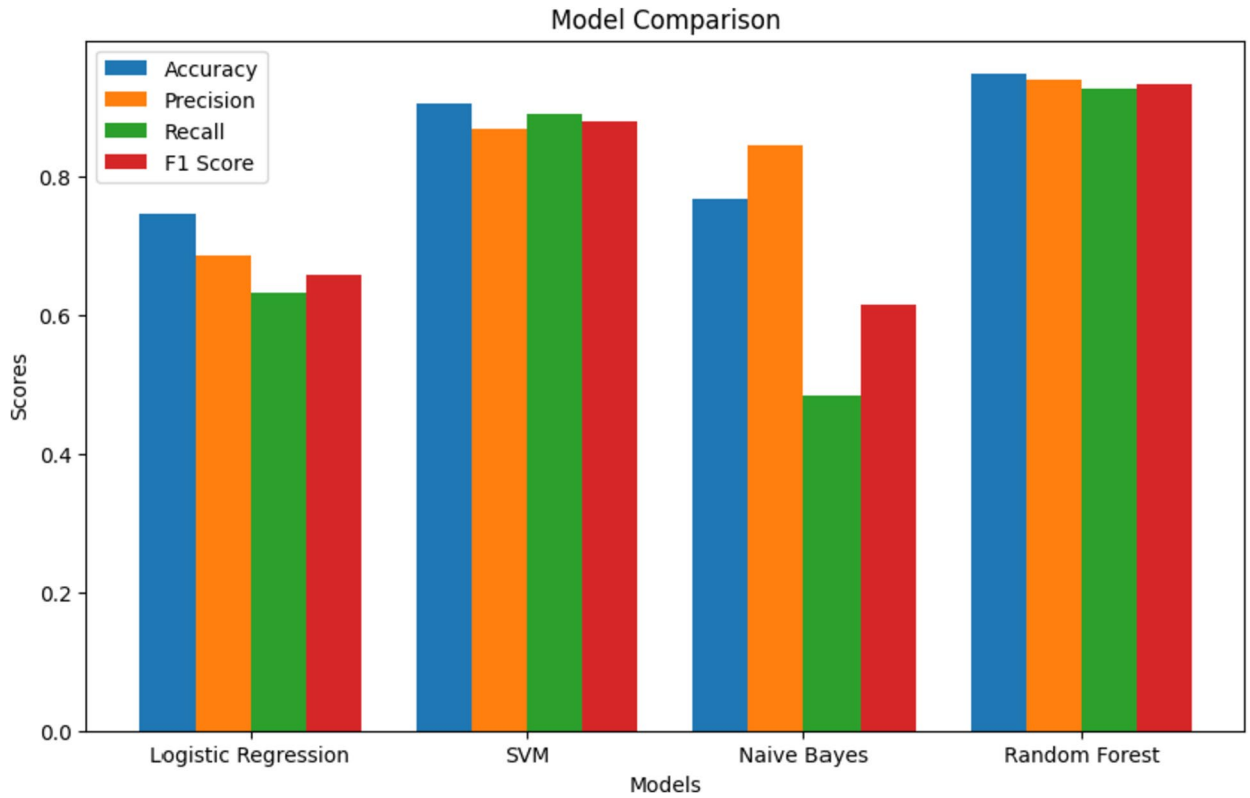

Fig. 3 Performance comparison of different models

noticeable in one of the metrics where it has the lowest bar among all models, indicating a weaker performance in that area.

The Naive Bayes model shows a mixed performance with one metric having a notably lower bar compared to the other models, suggesting it might be less reliable in that aspect of aggression detection.

The exact performance numbers for each metric are not visible in the chart, but the relative heights of the bars provide a visual comparison of the model performances. The graph helps to assess which model might be the most effective for implementing a cyber troll detection system, considering the balance between false positives, false negatives, and correctly identified instances. Based on this visual representation, the Random Forest model would likely be the first choice for further validation and potential deployment.

Figure 4 displays Receiver Operating Characteristic (ROC) curves for five different machine learning models: Random Forest, LightGBM (Light Gradient Boosting Machine), Logistic Regression, SVM (Support Vector Machine), and Naive Bayes. The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

The area under the ROC curve (AUC) is a measure of the model's ability to distinguish between the classes and is generally considered as one of the most important evaluation metrics for checking any classification model's performance. A model with an AUC closer to 1 indicates better performance, whereas an AUC closer to 0.5 suggests no discriminative ability better than random chance.

From the provided image, we can infer the following about the performance of the models:

Random forest

The ROC curve is almost a 45-degree line, which is indicative of a model with no classification capability ($AUC \approx 0.49$). This suggests that the Random Forest model is not performing well in distinguishing between the positive and negative classes for this specific task.

LightGBM

The curve hugs the top left corner, indicating a high true positive rate and a low false positive rate, which is desirable in a good classifier. The AUC is very high ($AUC \approx 0.95$), showing excellent performance.

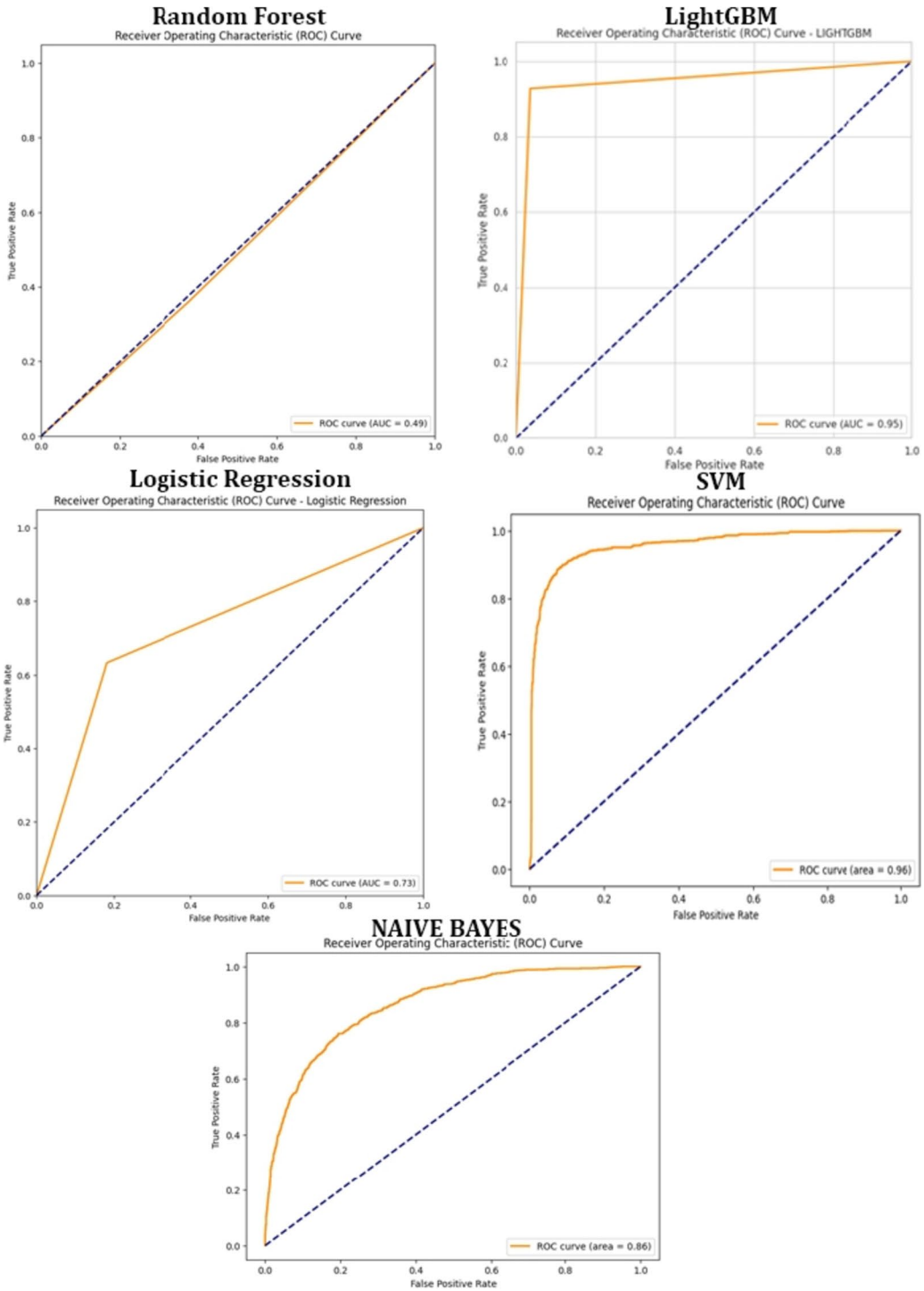

Fig. 4 Performance comparison of different models using ROC Curve

Logistic regression

The ROC curve shows a moderate performance with an AUC of around 0.73, suggesting it has a reasonable ability to distinguish between the classes, although not as effectively as LightGBM.

SVM

The ROC curve for the SVM is very close to the top left corner, similar to LightGBM, indicating a very high AUC (AUC $\approx$ 0.96), which means the SVM has an excellent discrimination capacity for the given classification task.

Naive bayes

This model's ROC curve is above the line of no-discrimination, with an AUC of about 0.85, suggesting it has a good performance, although not as strong as LightGBM or SVM.

In summary, based on the ROC curves, SVM and LightGBM are the top-performing models for this particular classification problem, followed by Naive Bayes and Logistic Regression, with Random Forest performing poorly. It is important to note that these curves are useful for visualizing and comparing the performance of different models but should be complemented with other metrics and analyses to fully understand model performance in practical applications.

Deep learning finds diverse applications across the reviewed studies, showcasing its versatility and significance in various domains. In Yu et al.'s research (2021), deep learning can be applied for anomaly detection to enhance security in touch screen devices, helping identify and prevent indirect eavesdropping attacks [28]. In the field of LiDAR data processing, as presented by Zhou et al. (2021), deep learning can be leveraged for efficient signal decomposition, contributing to improved LiDAR data analysis and interpretation [29]. Qi et al.'s work (2022) on brightness correction offers opportunities for deep learning-based image enhancement and quality improvement, particularly in multi-region nonuniform scenarios [30]. Cao et al. (2021) propose reliable communication in wireless-powered NOMA systems, where deep learning can optimize resource allocation and enhance system performance [31].

Furthermore, Wu et al.'s study (2022) on dynamic spectrum allocation in cognitive radio networks suggests that deep learning can optimize pricing policies and resource allocation, improving spectrum utilization efficiency [32]. Li et al. (2022) introduce smartphone app usage analysis, where deep learning can be employed for behavior pattern recognition and user profiling, aiding app developers and marketers [33]. In the context of adaptive co-site interference cancellation, Jiang and Li (2022) indicate the potential of deep learning in interference mitigation and signal processing [34]. Deep learning's applications extend to the educational domain, with Huang et al. (2021) proposing sentiment analysis and interaction level assessment using learning analytics, aiding in understanding and improving blended learning environments [35]. In spam detection, Wu et al.'s hybrid PU-learning-based model (2020) can benefit from deep learning techniques to enhance the accuracy and efficiency of spammer detection [36].

Li et al. (2023) explore public-key authenticated encryption with keyword search, which can leverage deep learning for fast and accurate search operations in encrypted data [37]. Sun et al.'s work (2020) on low-latency service function chaining orchestration in network function virtualization can employ deep learning for efficient decision-making and orchestration of network functions [38]. Similarly, Sun et al. (2019) and Sun et al. (2018) demonstrate cost-efficient and domain-spanning service function chain orchestration, where deep learning can optimize service placement and chaining decisions across multiple domains [39, 40]. Li et al. (2022) investigate daily activity patterns in smartphone app usage, presenting an opportunity for deep learning to identify and predict user behaviors, enhancing user experiences and app recommendations [41]. Furthermore, Liu et al. (2023) propose Sketch2Photo, which can benefit from deep learning techniques to improve the synthesis of photo-realistic images from sketches, enabling various creative applications [42]. In the context of developing multi-labeled corpora for Twitter short texts, Liu et al. (2023) illustrate how deep learning can assist in text analysis and classification [43]. Li et al. (2023) explore the computational effects of advanced deep neural networks on logical and activity learning, emphasizing the role of deep learning in enhancing cognitive skills and thinking processes [44]. Lastly, Zhang et al. (2023) present a security defense decision method for complex networks, where deep learning can be employed for anomaly detection and threat identification, contributing to network security [45].

The study's practical implications are significant in the context of addressing cyber-trolling behaviors and enhancing online safety. Firstly, the finding that the Random Forest classifier outperformed other models in detecting cyber troll posts underscores the importance of employing ensemble methods and robust algorithms when developing automated tools for aggression detection in digital communications. Organizations and online platforms seeking to implement troll detection systems can benefit from adopting Random Forest-based approaches, as they demonstrate superior accuracy and a balanced trade-off between precision and recall, which is crucial for minimizing false positives and false negatives

in identifying cyber trolls. Secondly, the observation that LightGBM tended higher false predictions suggests that while gradient boosting algorithms can be effective, careful parameter tuning and model evaluation are essential to mitigate false positives and ensure the reliability of detection systems. This insight guides practitioners in the selection and optimization of machine learning models tailored for cyber troll detection.

The anomaly identified among Logistic Regression, SVM, and Naive Bayes classifiers raises concerns about their suitability for this specific task [46–48]. The practical implication here is the need for meticulous data preprocessing and feature engineering, as well as a rigorous model assessment when using these algorithms for text classification in social media contexts. Future research and development efforts should focus on understanding the reasons behind this anomaly and refining the application of these classifiers for cyber troll detection. Furthermore, the study emphasizes the importance of transparency and interpretability in machine-learning models designed for online safety. Cyber troll detection systems must not only perform effectively but also provide interpretable results, enabling human moderators and administrators to understand and act upon the model's predictions. This underscores the need for further research into explainable AI techniques and their integration into the development of troll detection tools. Lastly, the mention of future work involving deep learning techniques hints at the potential for further advancements in cyber troll detection. Deep learning models, such as recurrent neural networks (RNNs) and transformer-based architectures, have shown promise in natural language processing tasks and may offer improved performance in this domain. The study encourages future investigations into the applicability of these advanced techniques and their ability to enhance cyber troll detection accuracy.

Conclusion

The present study has provided valuable insights into the effectiveness of various machine learning classifiers in the context of detecting cyber-trolling behaviors in digital communications. Through a rigorous evaluation of Random Forest, Light Gradient Boosting Machine (LightGBM), Logistic Regression, Support Vector Machine (SVM), and Naive Bayes classifiers on a publicly available dataset, we have uncovered practical implications for enhancing online safety. In conclusion, the Random Forest classifier emerged as the top-performing model, showcasing the highest accuracy and achieving a balanced precision-recall trade-off. This finding underscores the significance of employing ensemble methods when developing automated tools for identifying cyber trolls. However, it is essential to emphasize that while Random Forest exhibited superior performance, other classifiers like LightGBM also demonstrated efficacy, albeit with some tendency towards higher false predictions. This suggests that gradient boosting algorithms can be effective but require careful parameter tuning and model evaluation. The anomaly observed among Logistic Regression, SVM, and Naive Bayes classifiers highlights the need for cautious data preprocessing and feature engineering when applying these algorithms in the realm of cyber troll detection. Further investigation is warranted to understand the reasons behind this anomaly and to optimize the application of these classifiers.

Future work

Building on the findings of this study, several avenues for future research and development in the field of cyber troll detection can be identified:

Anomaly investigation

Further exploration into the anomaly observed among Logistic Regression, SVM, and Naive Bayes classifiers is imperative. This entails a detailed examination of data characteristics, feature extraction methods, and potential limitations in the model application process. Identifying and addressing these issues can lead to improved performance and a better understanding of the suitability of these algorithms for cyber troll detection.

Deep learning approaches

As alluded to in the study, the potential of deep learning techniques, including recurrent neural networks (RNNs) and transformer-based models, should be explored. These advanced architectures have demonstrated remarkable capabilities in natural language processing tasks and may offer enhanced performance in detecting nuanced forms of cyber trolling.

Explainable AI

Ensuring transparency and interpretability in model predictions is crucial, particularly for online safety systems. Future work should delve into the integration of explainable AI techniques to enable human moderators and administrators to comprehend and trust the model's decisions. This is especially important in a context where action needs to be taken based on the model's output.

Real-time implementation

Developing real-time cyber troll detection systems that can be seamlessly integrated into various online platforms and social media networks is a pressing need. Future research should focus on the scalability and

efficiency of detection algorithms to handle large volumes of digital communications in real-time.

Cross-domain generalization

Investigating the generalization of the developed models across different online platforms and linguistic domains is essential. The robustness and adaptability of the models should be assessed to ensure their effectiveness in diverse online environments.

In conclusion, this study lays the foundation for further advancements in the field of cyber troll detection. Future research endeavors should address the identified anomalies, explore deep learning approaches, prioritize explainable AI, work towards real-time implementation, and assess cross-domain generalization to continue the pursuit of a safer and more inclusive digital space.

Authors' contributions

Huimin Han: Conceptualization, Methodology, Writing - Original Draft. Muhammad Asif: Methodology, Data Analysis, Writing - Review & Editing. Emad Mahrous Awwad: Data Collection, Validation, Writing- Review & Editing. Nadia Sarhan: Supervision, Funding Acquisition, Writing- Review & Editing. Yazeed Yasid Ghadi: Data Analysis, Visualization, Writing - Review & Editing. Bo Xu: Conceptualization, Supervision, Writing- Review & Editing.

Funding

The authors present their appreciation to King Saud University for funding this research through the Researchers Supporting Program number (RSPD2024R1052), King Saud University, Riyadh, Saudi Arabia and Hainan University Startup Fund KYQD(ZR)23143.

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1]Mechanical and Electrical Engineering College, Hainan Vocational University of Science and Technology, Haikou 571126, China. [2]School of Media, Hunan University of Science and Engineering, Yongzhou 425199, Hunan, China. [3]Department of Electrical Engineering, College of Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia. [4]Department of Quantitative Analysis, College of Business Administration, King Saud University, Riyadh, Saudi Arabia. [5]Department of Computer Science, Al Ain University, Al Ain, UAE. [6]School of information and communication Engineering, Hainan University, Hainan, China.

Received: 6 December 2023 Accepted: 19 December 2023
Published online: 16 January 2024

References

1. Garett R, Lord LR, Young SD (2016) Associations between social media and cyberbullying: a review of the literature. Mhealth 2:46. https://doi.org/10.21037/mhealth.2016.12.01
2. Selkie EM, Kota R, Moreno M, CYBERBULLYING BEHAVIORS AMONG FEMALE, Coll Stud J (2016) Spring ;50(2):278–287
3. Leung ANM (2021) To help or not to help: intervening in Cyberbullying among Chinese Cyber-bystanders. Front Psychol 12:483250. https://doi.org/10.3389/fpsyg.2021.483250
4. Doumas DM, Midgett A (2020) Witnessing cyberbullying and internalizing symptoms among Middle School Students. Eur J Investig Health Psychol Educ 10(4):957–966. https://doi.org/10.3390/ejihpe10040068
5. Zhan J, Yang Y, Lian R (2022) The relationship between cyberbullying victimization and cyberbullying perpetration: the role of social responsibility. Front Psychiatry 13:995937. https://doi.org/10.3389/fpsyt.2022.995937
6. Lam TN, Jensen DB, Hovey JD, Roley-Roberts ME (2022) College students and cyberbullying: how social media use affects social anxiety and social comparison. Heliyon 8(12):e12556. https://doi.org/10.1016/j.heliyon.2022.e12556
7. Selkie EM, Kota R, Chan YF, Moreno M (2015) Cyberbullying, depression, and problem alcohol use in female college students: a multisite study. Cyberpsychol Behav Soc Netw 18(2):79–86. https://doi.org/10.1089/cyber.2014.0371
8. Chanda SS, Banerjee DN (2022) Omission and commission errors underlying AI failures. AI & Soc. https://doi.org/10.1007/s00146-022-01585-x
9. Nizamani AH, Chen Z, Nizamani AA, Bhatti UA (2023) Advance Brain Tumor segmentation using feature fusion methods with deep U-Net model with CNN for MRI data. J King Saud University-Computer Inform Sci 35(9):101793
10. Zhang Y, Chen J, Ma X, Wang G, Bhatti UA, Huang M (2024) Interactive medical image annotation using improved attention U-net with compound geodesic distance. Expert Syst Appl 237:121282
11. Chen Y, Zhou Y, Zhu S, Xu H (2012) Detecting Offensive Language in Social Media to protect adolescent online safety. 2012 Int Conf Priv Secur Risk Trust 2012 Int Confernece Social Comput Amsterdam Neth 71–80. https://doi.org/10.1109/SocialCom-PASSAT.2012.55
12. Gaydhani A, Doma V, Kendre, Shrikant, Laxmi BB (2018) Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach
13. Yin W, Zubiaga A (2022) Hidden behind the obvious: misleading keywords and implicitly abusive language on social media. Online Social Networks and Media Volume 30 100210:2468–6964. https://doi.org/10.1016/j.osnem.2022.100210
14. Bohr A, Memarzadeh K (2020) The rise of artificial intelligence in healthcare applications. Artif Intell Healthc 25–60. https://doi.org/10.1016/B978-0-12-818438-7.00002-2
15. Taherdoost H (2023) Enhancing Social Media Platforms with Machine Learning algorithms and neural networks. Algorithms 16:271. https://doi.org/10.3390/a16060271
16. Conway M, Hu M, Chapman WW (2019) Recent advances in Using Natural Language Processing To Address Public Health Research Questions Using Social Media and ConsumerGenerated Data. Yearb Med Inform 28(1):208–217. https://doi.org/10.1055/s-0039-1677918Epub 2019 Aug 16
17. Agathe Balayn J, Yang Z, Szlavik, Bozzon A (2021) Automatic Identification of Harmful, Aggressive, Abusive, and Offensive Language on the Web: A Survey of Technical Biases Informed by Psychology Literature. Trans. Soc. Comput. 4, 3, Article 11 (September 2021), 56 pages. https://doi.org/10.1145/3479158
18. Alrashidi B, Jamal A, Khan I, Alkhathlan A (2022) A review on abusive content automatic detection: approaches, challenges and opportunities. PeerJ Comput Sci 8:e1142. https://doi.org/10.7717/peerj-cs.1142
19. Nascimento FRS, Cavalcanti GDC, Da Costa-Abreu M (2023) Exploring Automatic hate Speech Detection on Social Media: a focus on content-based analysis. SAGE Open 13(2). https://doi.org/10.1177/21582440231181311
20. Bhatti UA, Tang H, Wu G, Marjan S, Hussain A (2023) Deep learning with graph convolutional networks: an overview and latest applications in computational intelligence. Int J Intell Syst 2023:1–28
21. Bhatti UA, Huang M, Neira-Molina H, Marjan S, Baryalai M, Tang H, … Bazai, S. U. (2023) MFFCG–Multi feature fusion for hyperspectral image classification using graph attention network. Expert Syst App 229:120496
22. Shickel B, Tighe PJ, Bihorac A, Rashidi P, Deep EHR (2018) A survey of recent advances in Deep Learning Techniques for Electronic Health Record (EHR) analysis. IEEE J Biomed Health Inform 22(5):1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

23. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, DeVylder J, Walter M, Berrouiguet S, Lemey C (2021) Machine Learning and Natural Language Processing in Mental Health: systematic review. J Med Internet Res 23(5):e15708. https://doi.org/10.2196/15708
24. Pennacchiotti M, Popescu A (2011) A Machine Learning Approach to Twitter User Classification. *Proceedings of the International AAAI Conference on Web and Social Media*
25. Sarwar SM, Murdock V (2021) Unsupervised Domain Adaptation for Hate Speech Detection Using a Data Augmentation Approach
26. Chen ZS (2022) Prathamesh (Param) Kulkarni, Isaac R. Galatzer-Levy, Benedetta Bigio, Carla Nasca, Yu Zhang. Modern views of machine learning for precision psychiatry. Patterns, Volume 3, Issue 11, 100602, ISSN 2666–3899, https://doi.org/10.1016/j.patter.2022.100602
27. Muneer A, Fati SM (2020) A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. Future Internet 12:187. https://doi.org/10.3390/fi12110187
28. Yu J, Lu L, Chen Y, Zhu Y, Kong L (2021) An indirect eavesdropping Attack of keystrokes on Touch screen through Acoustic Sensing. IEEE Trans Mob Comput 20(2):337–351. https://doi.org/10.1109/TMC.2019.2947468
29. Zhou, G., Deng, R., Zhou, X., Long, S., Li, W., Lin, G.,… Li, X. (2021). Gaussian Inflection Point Selection for LiDAR Hidden Echo Signal Decomposition. IEEE geoscience and remote sensing letters, 1–5. doi: https://doi.org/10.1109/LGRS.2021.3107438
30. Qi, M., Cui, S., Chang, X., Xu, Y., Meng, H., Wang, Y.,… Arif, M. (2022). Multi-region Nonuniform Brightness Correction Algorithm Based on L-Channel Gamma Transform. Security and communication networks, 2022. doi: https://doi.org/10.1155/2022/2675950
31. Cao, K., Wang, B., Ding, H., Lv, L., Tian, J., Hu, H.,… Gong, F. (2021). Achieving Reliable and Secure Communications in Wireless-Powered NOMA Systems. IEEE transactions on vehicular technology, 70(2), 1978–1983. doi: https://doi.org/10.1109/TVT.2021.3053093
32. Wu H, Jin S, Yue W (2022) Pricing policy for a dynamic Spectrum Allocation Scheme with batch requests and impatient packets in Cognitive Radio Networks. J Syst Sci Syst Eng 31(2):133–149. https://doi.org/10.1007/s11518-022-5521-0
33. Li, T., Xia, T., Wang, H., Tu, Z., Tarkoma, S., Han, Z.,… Hui, P. (2022). Smartphone App Usage Analysis: Datasets, Methods, and Applications. IEEE Communications Surveys& Tutorials, 24(2), 937–966. doi: https://doi.org/10.1109/COMST.2022.3163176
34. Jiang Y, Li X (2022) Broadband cancellation method in an adaptive co-site interference cancellation system. Int J Electron 109(5):854–874. https://doi.org/10.1080/00207217.2021.1941295
35. Huang C, Han Z, Li M, Wang X, Zhao W (2021) Sentiment evolution with interaction levels in blended learning environments: using learning analytics and epistemic network analysis. Australasian J Educational Technol 37(2):81–95. https://doi.org/10.14742/ajet.6749
36. Wu, Z., Cao, J., Wang, Y., Wang, Y., Zhang, L.,… Wu, J. (2020). hPSD: A Hybrid PU-Learning-Based Spammer Detection Model for Product Reviews. IEEE transactions on cybernetics, 50(4),1595–1606. doi: https://doi.org/10.1109/TCYB.2018.2877161
37. Li H, Huang Q, Huang J, Susilo W (2023) Public-key authenticated encryption with Keyword Search supporting constant Trapdoor Generation and fast search. IEEE Trans Inf Forensics Secur 18:396–410. https://doi.org/10.1109/TIFS.2022.3224308
38. Sun, G., Xu, Z., Yu, H., Chen, X., Chang, V.,… Vasilakos, A. V. (2020). Low-Latency and Resource-Efficient Service Function Chaining Orchestration in Network Function Virtualization. IEEE Internet of Things Journal, 7(7), 5760–5772. doi: https://doi.org/10.1109/JIOT.2019.2937110
39. Sun, G., Zhu, G., Liao, D., Yu, H., Du, X.,… Guizani, M. (2019). Cost-Efficient Service Function Chain Orchestration for Low-Latency Applications in NFV Networks. IEEE Systems Journal, 13(4), 3877–3888. doi: https://doi.org/10.1109/JSYST.2018.2879883
40. Sun G, Li Y, Liao D, Chang V (2018) Service function chain Orchestration Across multiple domains: a full mesh Aggregation Approach. IEEE Trans Netw Serv Manage 15(3):1175–1191. https://doi.org/10.1109/TNSM.2018.2861717
41. Li, T., Li, Y., Hoque, M. A., Xia, T., Tarkoma, S.,… Hui, P. (2022). To What Extent We Repeat Ourselves? Discovering Daily Activity Patterns Across Mobile App Usage.IEEE Transactions on Mobile Computing, 21(4), 1492–1507. doi: https://doi.org/10.1109/TMC.2020.3021987
42. Liu H, Xu Y, Chen F (2023) Sketch2Photo: synthesizing photo-realistic images from sketches via global contexts. Eng Appl Artif Intell 117:105608. https://doi.org/10.1016/j.engappai.2022.105608
43. Liu, X., Zhou, G., Kong, M., Yin, Z., Li, X., Yin, L.,… Zheng, W. (2023). Developing Multi-Labelled Corpus of Twitter Short Texts: A Semi-Automatic Method. Systems, 11(8),390. doi: https://doi.org/10.3390/systems11080390
44. Li D, Ortegas KD, White M (2023) Exploring the computational effects of Advanced Deep neural networks on logical and activity learning for enhanced thinking skills. Systems 11(7):319. https://doi.org/10.3390/systems11070319
45. Zhang, H., Mi, Y., Fu, Y., Liu, X., Zhang, Y., Wang, J.,… Tan, J. (2023). Security defense decision method based on potential differential game for complex networks.Computers & Security, 129, 103187. https://doi.org/10.1016/j.cose.2023.103187
46. Qasim M, Khan M, Mehmood W, Sobieczky F, Pichler M, Moser B (2022) A Comparative Analysis of Anomaly Detection Methods for Predictive Maintenance in SME. In:, et al. Database and Expert systems Applications - DEXA 2022 Workshops. DEXA 2022. Communications in Computer and Information Science, vol 1633. Springer, Cham. https://doi.org/10.1007/978-3-031-14343-4_3
47. Khan M, Liu M, Dou W, Yu S vGraph: Graph Virtualization towards Big Data, 2015 Third International Conference on Advanced Cloud and Big Data, 2015, pp. 153–158, https://doi.org/10.1109/CBD.2015.33
48. Rafique W, Khan M, Sarwar N, Sohail M, Irshad A (2019) A Graph Theory based method to Extract Social structure in the Society. In: Bajwa I, Kamareddine F, Costa A (eds) Intelligent Technologies and Applications. INTAP 2018. Communications in Computer and Information Science, vol 932. Springer, Singapore. https://doi.org/10.1007/978-981-13-6052-7_38

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.