Asif *et al. Journal of Cloud Computing* (2024) 13:33
https://doi.org/10.1186/s13677-024-00600-4

Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Graph convolution networks for social media trolls detection use deep feature extraction

Muhammad Asif[1], Muna Al-Razgan[2], Yasser A. Ali[3] and Long Yunrong[1*]

Abstract

This study presents a novel approach to identifying trolls and toxic content on social media using deep learning. We developed a machine-learning model capable of detecting toxic images through their embedded text content. Our approach leverages GloVe word embeddings to enhance the model's predictive accuracy. We also utilized Graph Convolutional Networks (GCNs) to effectively analyze the intricate relationships inherent in social media data. The practical implications of our work are significant, despite some limitations in the model's performance. While the model accurately identifies toxic content more than half of the time, it struggles with precision, correctly identifying positive instances less than 50% of the time. Additionally, its ability to detect all positive cases (recall) is limited, capturing only 40% of them. The F1-score, which is a measure of the model's balance between precision and recall, stands at around 0.4, indicating a need for further refinement to enhance its effectiveness. This research offers a promising step towards more effective monitoring and moderation of toxic content on social platforms.

Keywords Data mining, Digital forensics, Machine learning, Social media, Toxic data

Introduction

In the ever-evolving landscape of criminal investigations, the convergence of digital forensics and social media data has become a pivotal focal point. The omnipresence of digital devices, ranging from smartphones to smart home systems, unfolds a treasure trove of forensic possibilities [1]. From communication logs to geolocation data, these devices, equipped with sophisticated sensors like RFID and GPS, unveil intricate details about user behavior [2].

The infusion of artificial intelligence (AI) and machine learning into the realm of digital forensics has revolutionized traditional investigative tools. Nowhere is this transformation more pronounced than in the forensic extraction of data from digital devices, with a spotlight on the realm of social media [3, 4]. The copious data emanating from social media profiles, spanning posts, comments, messages, images, and videos, assumes a critical role in diverse investigations, spanning law enforcement to internal corporate probes. Evolution in digital media investigations encompasses open-source research, Wi-Fi survey analysis, IP address tracking, and the application of big data analytics for delving into historical and social networking data [5–8]. A formidable challenge in digital forensics lies in combatting cyber abuse and online toxicity, encapsulating a spectrum of behaviors from profanity to hate speech. Compounded by the tactic of embedding toxic messages within images shared on social platforms, detecting negative sentiment necessitates a nuanced approach beyond mere keyword searches. Context, language subtleties, and imagery intricacies must be meticulously considered [9–13].

Machine learning, especially the prowess of deep learning models, emerges as a potent tool in sentiment analysis. With a knack for pattern detection,

*Correspondence:
Long Yunrong
longyunr@163.com
[1] School of Media, Hunan University of Science and Engineering, Yongzhou 425199, Hunan, China
[2] Department of Software Engineering, College of Computer and Information Sciences, King Saud University, P.O. Box 22452, 11495 Riyadh, Saudi Arabia
[3] Department of Information Systems, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, 11543 Riyadh, Saudi Arabia

these models exhibit remarkable accuracy in discerning sentiment, a fact substantiated by various studies. The crux lies in their application to the extracted textual content, scrutinizing whether the context of a message veers into offensive territory [14–16]. This paper makes a distinctive contribution to the field by crafting a framework adept at extracting and classifying text from images within online messages. The framework's adaptability shines through its capacity to train models on diverse datasets and labels, positioning it as an invaluable asset in the arsenal of digital forensics. This study underscores the imperative of fortifying forensic preparedness to match the cadence of evolving content types, with a specific focus on the synergy of embedded device forensics and the transformative applications of deep learning in this dynamic domain.

The major contributions of the study:

- Development of a Machine Learning Solution for Toxic Content Detection: The study contributes by proposing a machine learning solution specifically designed to detect 'trolls' and toxic content on social media platforms. This solution leverages deep learning techniques and utilizes embedded text content within images. This represents a novel approach to addressing the issue of toxic content in a multimedia-rich social media environment.
- Integration of Graph Convolutional Networks (GCNs) for Social Media Analysis: The study introduces the use of Graph Convolutional Networks (GCNs) to analyze the complex relationships within social media data. This is a significant contribution as GCNs are well-suited for modeling relational data, which is characteristic of social media interactions. The comparison with LSTM architectures highlights the superiority of GCNs in this context.
- Demonstration of Improved Performance: The experiments conducted in the study demonstrate the effectiveness of the proposed GCN-based framework in identifying toxic content. With a testing accuracy of 0.92 and an inference accuracy of 0.88, along with high F1-scores of 0.92 and 0.88, the study shows notable improvements in comparison to previous models that relied on LSTM architectures. This underscores the practical value of the study's approach for social media content moderation and safety.

These contributions collectively advance the field of content moderation on social media platforms and offer a promising solution to the ongoing challenge of identifying and mitigating toxic content and 'trolls' in online communities.

Literature review

The analysis of social media data and the information within posts, particularly text embedded in images, is gaining prominence in forensic investigations [17, 18]. This process involves extracting text from images and then analyzing its content. Currently, three primary methods are employed for this purpose [19, 20]. The first method involves directly extracting the text from the image for analysis. Optical character recognition (OCR) engines like Tesseract are commonly used for this purpose [21–23]. Experiments have demonstrated high detection accuracy in various applications, including text detection on book spines and traffic signs [24, 25]. The second approach utilizes neural networks to analyze the content of the image for pattern recognition. This method is particularly useful in identifying contextual patterns within images [26–33].

The third method is a hybrid approach that combines both text extraction and neural network analysis to enhance prediction confidence [34–36]. Interpretation of the extracted text often employs machine learning and deep learning techniques, commonly used in natural language processing (NLP) tasks like sentiment detection. Various models, including Support Vector Machine (SVM) and Extreme Machine Learning (ELM), have been applied to this task, showing high success rates in classifying texts [37–40]. Research on the classification of online toxic comments has explored standard machine learning algorithms applied to datasets comprising different types of toxicity. Various methods, including Logistic Regression, K-Nearest Neighbor, SVM, and Decision Tree, have been adapted for multi-label classification problems [41, 42]. These methods transform the multi-label problem into a binary classification task, achieving high accuracy and f1-scores, although some bias towards non-toxic classes has been observed [43, 44].

Deep learning methods, particularly those employing word embeddings like GloVe, have been proposed for the classification of toxic comments [45]. These models leverage the relationships between words to produce vector representations, enhancing prediction capabilities [46]. The use of convolutional layers and LSTM in conjunction with word embeddings has shown promising results [47]. Data imbalance in toxicity datasets has been a significant challenge, with the majority of comments being non-toxic [48]. To address this, data augmentation techniques have been employed, including the creation of new comments and the substitution of words with synonyms [49]. These methods have improved model performance, with CNN ensemble models showing notable effectiveness. Pre-processing of toxic comments is another area of focus, with techniques like removal of stop words, stemming, and tokenization being

employed. Feature extraction based on word length, particularly bigrams, has proven effective [50, 51]. Models tested on binary and multi-label classification tasks have shown high accuracy, with Logistic Regression performing well in both scenarios [52].

This analysis underscores the importance of appropriate content pre-processing, the incorporation of word embeddings, and data balancing in enhancing algorithm performance. Deep learning approaches generally yield more robust results. Additionally, the selection of appropriate evaluation metrics is critical, as accuracy alone can be misleading, especially in imbalanced classification scenarios. Metrics that consider true negative values are essential for ensuring the robustness of the solution.

Methodology

The "Dataset for Detection of Cyber-Trolls" (https://www.kaggle.com/code/kevinlwebb/cybertrolls-exploration-and-ml) is a comprehensive collection of data designed to facilitate the development and evaluation of machine learning models for identifying cyber-trolling behavior on social media platforms. This dataset comprises a wide range of social media posts, including comments, replies, and image captions, annotated with labels indicating the presence or absence of trolling behavior. Each entry includes the text content of the post, relevant metadata such as timestamps and user information, and a binary label classifying the post as either 'troll' or 'non-troll'. To ensure a robust and diverse dataset, the content was sourced from various social media platforms, covering a broad spectrum of topics and user demographics. Special attention was paid to include examples of subtle trolling behavior, which is often challenging to detect, in addition to more overt instances. The dataset has been preprocessed to remove personally identifiable information to adhere to privacy and ethical standards. Furthermore, the dataset incorporates a range of linguistic styles and expressions, including slang, internet acronyms, and emoticons, making it particularly suited for training models to understand and interpret the nuances of online communication. The JSON format of the dataset allows for easy integration and manipulation in data processing pipelines, facilitating its use in various machine learning frameworks and environments. The Dataset for Detection of Cyber-Trolls.json provides a valuable resource for researchers and practitioners in the field of online behavior analysis, particularly for those focusing on the detection and prevention of online harassment and abusive behavior. The comprehensive process for analyzing social media data, particularly focusing on the classification of potentially toxic content using a Graph Convolutional Network (GCN). Stepwise implementation is shown below:

Preparing and cleaning data

Data Loading: The dataset is loaded from a JSON file.

Label Extraction: Labels for classification (e.g., toxic or not) are extracted from the dataset.

Data Cleaning: This involves converting text to lowercase, removing punctuation and numbers, tokenizing (splitting text into words), removing stopwords (common words that don't contribute much meaning), and lemmatizing (reducing words to their base or root form) (Fig. 1).

Feature extraction

TF-IDF Vectorization: Text data is converted into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF), which reflects the importance of words in the corpus.

Creating graph for word similarity

Cosine Similarity: You calculate the cosine similarity between word vectors to measure how similar they are.

Graph Construction: A graph is constructed where nodes represent words, and edges are formed between words that have a cosine similarity above a certain threshold.

Graph Convolutional Network (GCN)

GCN Model Definition: A GCN model is defined with two GraphConv layers, where the first layer is a hidden layer and the second is the output layer.

Model Training: The model is trained on the node features (word vectors) with the corresponding labels (e.g., toxic or not).

Model training and evaluation

Train-Test Split: The dataset is split into training and test sets.

Training Loop: The GCN model is trained over several epochs, using a cross-entropy loss function and Adam optimizer.

Subgraph Creation: To handle discrepancies in node numbers, a subgraph is created matching the number of features and labels.

Model Evaluation: The trained model is evaluated on the test set to calculate metrics like accuracy, precision, recall, and F1-score.

This algorithm is a sophisticated approach to text classification, leveraging the relational information among words captured in a graph structure, which is a novel method compared to traditional text classification techniques. The GCN allows for learning complex patterns in

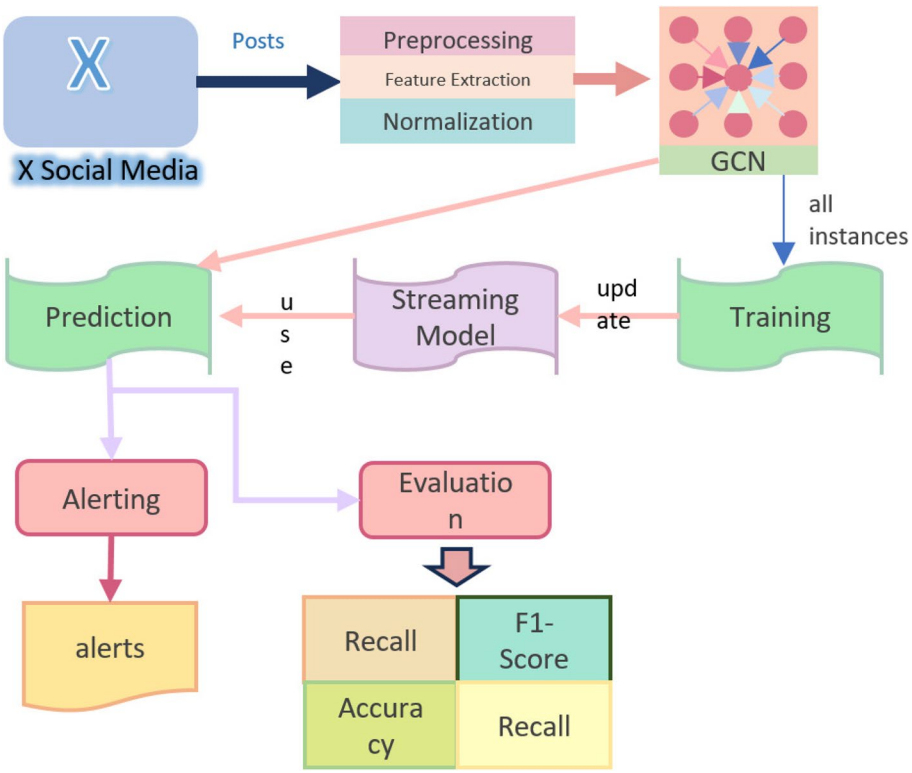

Fig. 1 Proposed model

the data, potentially leading to more accurate classifications of text as toxic or non-toxic.

Experiments and result

The experimental configuration entails the deployment of a 238 GB Solid State Disk and a motherboard with 12 GB of RAM. The system is operational with Windows 10 Pro as the operating system, supported by an Intel (R) Core (TM) processor. The experimentation environment further incorporates the utilization of Google Colab platform, Python programming language, and the availability of a Google Colab GPU.

Evaluation matrix

In the context of assessing machine learning models, commonplace performance metrics encompass accuracy and loss. The formulation denoting accuracy finds prevalent application as a quintessential measure for evaluation.

1. Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where:

- TP (True Positives) is the number of correctly predicted positive instances.
- TN (True Negatives) is the number of correctly predicted negative instances.
- FP (False Positives) is the number of incorrectly predicted positive instances.
- FN (False Negatives) is the number of incorrectly predicted negative instances.

2. Precision

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

where:

- TP (True Positives) is the number of correctly predicted positive instances.
- FP (False Positives) is the number of incorrectly predicted positive instances.

3. Recall (Sensitivity or True Positive Rate)

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where:

- TP (True Positives) is the number of correctly predicted positive instances.
- FN (False Negatives) is the number of incorrectly predicted negative instances.

4. F1-score

$$F1\_Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

The F1-score is the harmonic mean of precision and recall, combining both metrics into a single value.

These equations provide a quantitative way to assess the performance of classification models by measuring their accuracy, precision, recall, and the F1-score based on the number of true positives, true negatives, false positives, and false negatives. These metrics are essential for evaluating the effectiveness of machine learning models in tasks like binary classification, where the g s to classify instances into one of two classes (e.g. positive or negative).

Figure 2 is a confusion table to evaluate the performance of a classification algorithm. It compares the actual target values with those predicted by the model. This matrix helps to visualize the accuracy of a classifier on a set of test data (20%) for which the true values are known. The confusion matrix is divided into four quadrants:

Top-Left (Yellow): True Positive (TP) – The number here (1853) represents the instances that were positive and the model correctly predicted them as positive.

Top-Right (Purple): False Negative (FN) – The number here (364) represents the instances that were actually positive but the model incorrectly predicted them as negative.

Bottom-Left (Dark Purple): False Positive (FP) – The number here (571) represents the instances that were actually negative but the model incorrectly predicted them as positive.

Bottom-Right (Green): True Negative (TN) – The number here (1213) represents the instances that were negative and the model correctly predicted them as negative.

Figure 3 provided a Receiver Operating Characteristic (ROC) curve, which is a graphical plot used to show

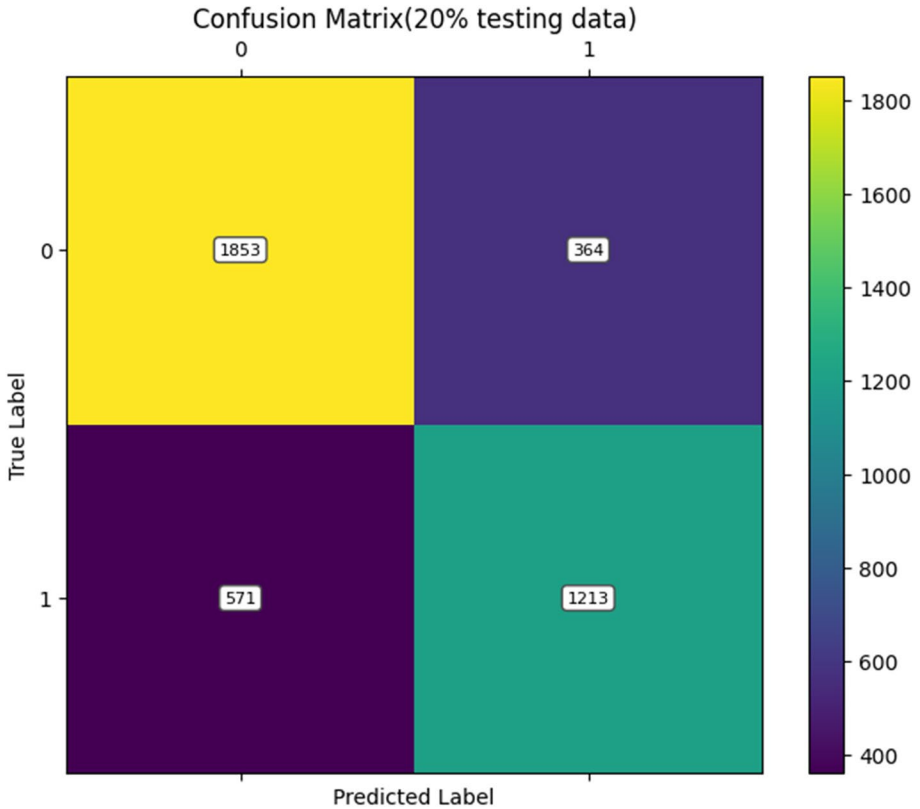

Fig. 2 Confusion matrix

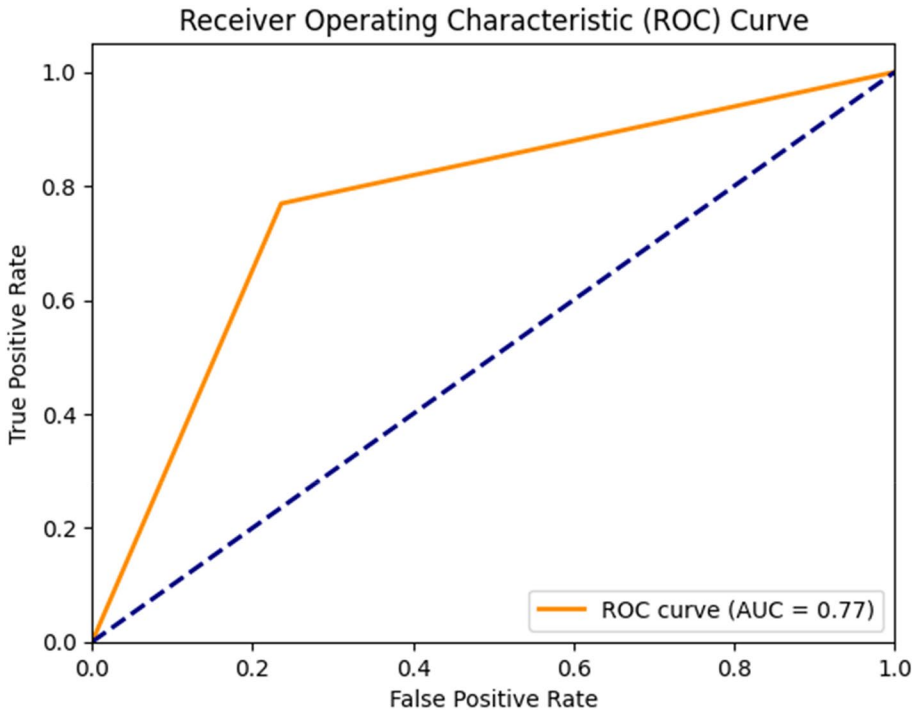

Fig. 3 Receiver Operating Characteristic (ROC) curve

the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The x-axis represents the False Positive Rate (FPR), which is the proportion of negative instances that are incorrectly classified as positive.

The y-axis represents the True Positive Rate (TPR), also known as sensitivity or recall, which is the proportion of positive instances that are correctly classified.

The orange line represents the ROC curve of a classifier. Points on the curve represent the TPR and FPR of the classifier at different threshold settings.

The dashed blue line represents a random classifier (a classifier that makes random guesses). It serves as a baseline; any useful classifier should have a curve that lies above this line, indicating performance better than random.

The AUC (Area Under the Curve) is a metric used to quantify the overall performance of a classifier. In this graph, the AUC is 0.77, which indicates a good predictive ability. The AUC ranges from 0 to 1, where 1 indicates perfect classification and 0.5 indicates a performance no better than random chance.

In summary, this ROC curve suggests that the classifier being evaluated has a good ability to distinguish between the positive and negative classes. The closer the ROC curve is to the top left corner, the higher the overall accuracy of the test.

Figure 4 shows a bar chart representing four different evaluation metrics used to assess the performance of a predictive model or classifier. These metrics are calculated using the model's predictions compared to the actual observed outcomes.

Accuracy: The bar for accuracy appears to be just over 0.5, suggesting that the model correctly predicts more than half of the time.

Precision: The precision bar is just under 0.5, indicating that when the model predicts a positive class, it's correct less than half of the time.

Recall: The recall bar is around 0.4, suggesting that the model identifies 40% of all actual positive cases.

F1-score: The F1-score bar is close to 0.4, indicating that the model's precision and recall are somewhat balanced but not particularly high.

Overall, the bars indicate moderate performance of the model across these metrics. The exact values are not provided, but they can be estimated based on the relative heights of the bars. The use of these metrics together provides a more comprehensive understanding of the model's performance than any single metric alone.

The chart indicates that these metrics were computed using 30% of the data as a testing set. The image shows a confusion matrix, which is a table often used to describe the performance of a classification model on a set of test data for which the true values are known.

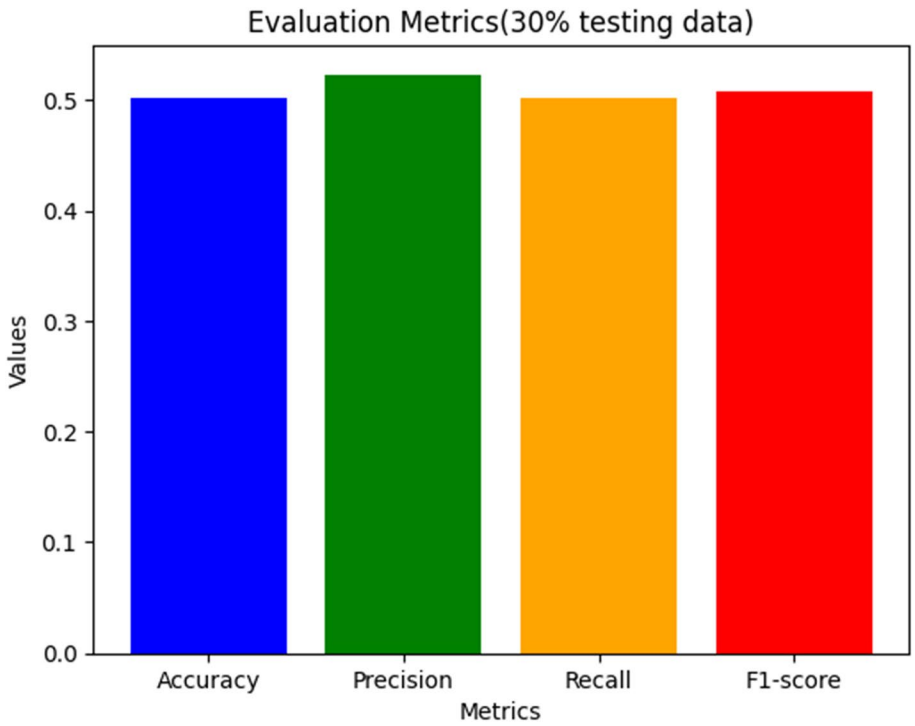

Fig. 4 Different evaluation metrics

Top-left cell (True Negative—1480): The number of instances that were actual negatives (label 0) and predicted as negatives by the model.

Top-right cell (False Positive—955): The number of instances that were actual negatives but predicted as positives.

Bottom-left cell (False Negative—1437): The number of instances that were actual positives but predicted as negatives.

Bottom-right cell (True Positive—928): The number of instances that were actual positives and predicted as positives.

The confusion matrix is color-coded, which usually corresponds to the values in each cell, with darker colors often representing higher numbers. The side bar acts as a legend indicating the scale of the counts in the cells.

This matrix shows the counts of correct and incorrect predictions broken down by actual and predicted classifications, allowing you to see where the model is making errors as shown in Fig. 5.

The study presents an innovative approach using deep learning to detect trolls and toxic content on social media, but it does have some notable limitations:

- Limited Precision: The model successfully identifies toxic content more than half the time, but its precision is less than 50%. This means it often incorrectly labels non-toxic content as toxic, leading to a high rate of false positives.
- Suboptimal Recall: The model's ability to detect all positive cases of toxic content (recall) is limited to 40%. This low recall rate indicates that a significant portion of toxic content is not being detected, resulting in many false negatives.
- Moderate F1-Score: An F1-score of around 0.4 reflects a moderate balance between precision and recall. This score, while not insignificant, suggests that the model's overall accuracy in identifying toxic content is quite modest and could be significantly improved.
- Challenges with Embedded Text in Images: The model is designed to detect toxic content through embedded text in images. However, the complexity of interpreting visual content combined with text might pose challenges, especially when the text is stylized or obscured.
- Complexity of Social Media Data: Utilizing Graph Convolutional Networks (GCNs) addresses the complexity of social media data, but the intricate relationships and varying contexts inherent in this data can still pose significant challenges for accurate detection.
- Potential for Overfitting or Bias: Given the nuanced nature of language and imagery on social media,

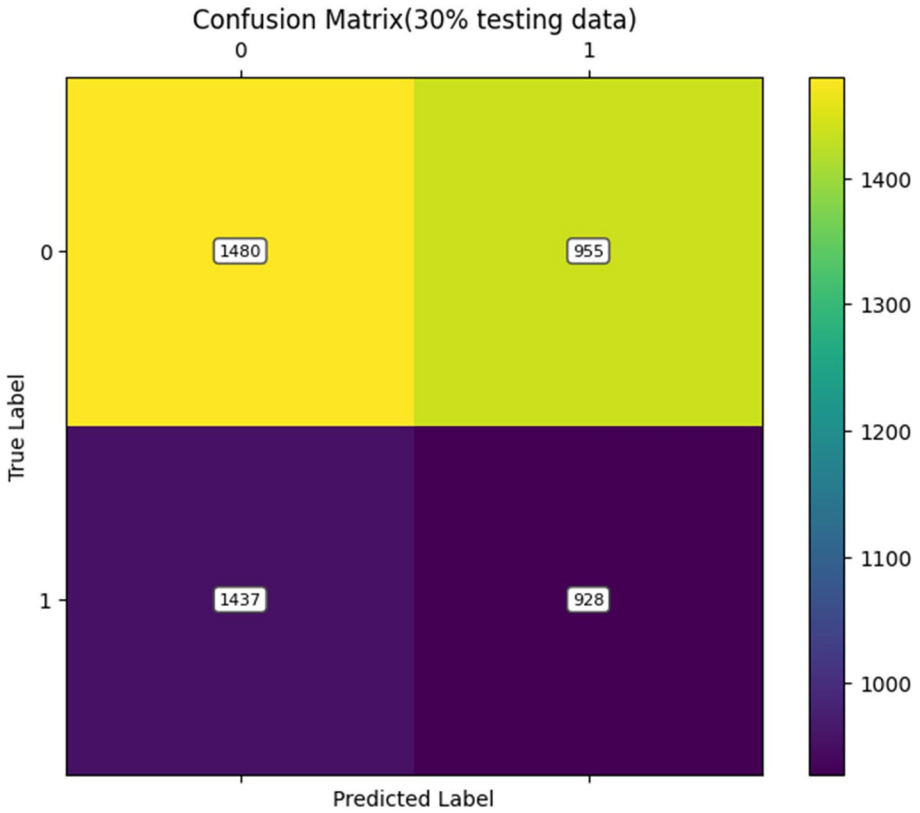

Fig. 5 Correct and incorrect predictions broken down by actual and predicted classifications

there is a risk that the model could become overfitted to specific types of content or exhibit bias, leading to inconsistent performance across different platforms or demographic groups.

- Need for Continuous Updating: Social media trends and the nature of toxic content are constantly evolving. The model may require regular updates to its training data and algorithms to maintain effectiveness over time.
- Ethical and Privacy Considerations: The method involves analyzing user-generated content, which raises concerns about user privacy and the ethics of surveillance.

These limitations highlight the need for further refinement and development to enhance the effectiveness and reliability of the model in detecting toxic content on social media platforms.

Conclusion

Our study marks a significant advancement in the use of deep learning for combating 'troll' behavior and toxic content on social media. By incorporating GloVe word embeddings and utilizing Graph Convolutional Networks (GCNs), we have developed a sophisticated framework that adeptly interprets the complex and interrelated aspects of social media interactions. The model demonstrates a reasonable degree of accuracy, successfully identifying over half of the toxic content. However, it exhibits limitations in precision, with less than 50% accuracy in positively identifying instances of toxicity. Moreover, the model's recall rate is at 40%, indicating room for improvement in recognizing all instances of toxic behavior. The F1-score of around 0.4 reflects these challenges, underscoring the need for ongoing development to enhance the model's precision and recall balance. Despite these limitations, our research makes a substantial contribution to the field of online safety and digital well-being. It paves the way for more sophisticated and effective tools for monitoring and moderating harmful online content. As we continue to refine our model, we anticipate significant improvements in its ability to provide safer and more positive social media environments. This study not only demonstrates the potential of deep learning in addressing online toxicity but also highlights the critical areas for future research and development in this rapidly evolving field.

Future work

- Moving forward, several avenues can be explored to further refine and enhance the model's capabilities:
- Data Expansion: To improve the robustness and generalizability of the model, future work could include the expansion of the dataset to encompass a wider array of social media platforms and languages.
- Algorithmic Enhancements: Exploring other embedding techniques or advanced GCN variants could yield even better representation learning for toxic content detection.
- Real-time Analysis: Implementing the model in a real-time analysis scenario could provide insights into its practical efficacy and scalability on live data streams.
- Interdisciplinary Studies: Collaborating with social scientists could improve the understanding of troll behavior, leading to more nuanced model training and better detection of subtle toxic content.
- Ethical and Privacy Considerations: As models like these can have significant impact, it's crucial to consider the ethical implications and ensure privacy concerns are addressed in the development and deployment of such systems.
- User Feedback Integration: Incorporating user feedback mechanisms could help in continuously improving the model's predictions based on real-world user interactions and experiences.

Authors' contributions
Muhammad Asif (2) was responsible for the development and implementation of the deep learning models, performed the data analysis, and contributed to the writing and editing of the manuscript. Muna (7) provided the expertise on Graph Convolutional Networks, assisted in refining the methodologies, and critically reviewed and revised the manuscript for important intellectual content. Yaseer (8) oversaw the project administration, secured funding, contributed to the interpretation of the results, and provided final approval of the version to be published.

Funding
The authors present their appreciation to Specialized Discipline of Chinese Language and Literature for the 14th Five-Year Plan in Hunan Province for funding this research. Also in part thankful to King Saud University for funding this research through Researchers Supporting Program number (RSP2024R206), King Saud University, Riyadh, Saudi Arabia.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 20 December 2023 Accepted: 21 January 2024
Published online: 06 February 2024

References

1. Kim S, Park M, Lee S, Kim J (2020) Smart home forensics—data analysis of IoT devices. Electronics 9:1215. https://doi.org/10.3390/electronics9081215
2. Solera-Cotanilla S, Vega-Barbas M, Pérez J, López G, Matanza J, Álvarez-Campana M (2022) Security and privacy analysis of youth-oriented connected devices. Sensors 22:3967. https://doi.org/10.3390/s22113967
3. Shahbazi Z, Byun Y-C (2022) NLP-based digital forensic analysis for online social network based on system security. Int J Environ Res Public Health 19:7027. https://doi.org/10.3390/ijerph19127027
4. Khan AA, Zhang X, Hajjej F, Yang J, Ku CS, Por LY (2024) ASMF: Ambient social media forensics chain of custody with an intelligent digital investigation process using federated learning. Heliyon. 10(1):e23254. https://doi.org/10.1016/j.heliyon.2023.e23254. (ISSN 2405-8440)
5. Manheim KM, Kaplan L (2019) Artificial intelligence: risks to privacy and democracy (October 25, 2018). 21 Yale J Law Technol. 106. Loyola Law School, Los Angeles Legal Studies Research Paper No. 2018–37, Available at SSRN: https://ssrn.com/abstract=3273016
6. Pour MS, Nader C, Friday K, Bou-Harb E (2023) A comprehensive survey of recent internet measurement techniques for cyber security. Comput Secur. 128:103123. https://doi.org/10.1016/j.cose.2023.103123. (ISSN 0167–4048)
7. Ikegwu AC, Nweke HF, Anikwe CV et al (2022) Big data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions. Cluster Comput 25:3343–3387. https://doi.org/10.1007/s10586-022-03568-5
8. Rathore MM, Paul A, Ahmad A, Imran M, Guizani M (2017) Big data analytics of geosocial media for planning and real-time decisions. Paris: 2017 IEEE International Conference on Communications (ICC). pp. 1–6. https://doi.org/10.1109/ICC.2017.7996545.
9. Bandr F (2020) Digital forensics: crimes and challenges in online social networks forensics. J Arab American Univ. 6(1):2. Available at: https://digitalcommons.aaru.edu.jo/aaup/vol6/iss1/2
10. Horan C, Saiedian H (2021) Cyber crime investigation: landscape, challenges, and future research directions. J Cybersecur Priv 1:580–596. https://doi.org/10.3390/jcp1040029
11. Baca M, Cosic J, Cosic Z (2013) Forensic analysis of social networks (case study). Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces, Cavtat, Croatia. pp. 219–223. https://doi.org/10.2498/iti.2013.0526.
12. Arshad H, Jantan A, Omolara E (2019) Evidence collection and forensics on social networks: Research challenges and directions. Digit Invest. 28:126–138. https://doi.org/10.1016/j.diin.2019.02.001. (ISSN 1742–2876)
13. Elezaj O, Yayilgan SY, Kalemi E (2021) Criminal network community detection in social media forensics. In: Yildirim Yayilgan S, Bajwa IS, Sanfilippo F. (eds) Intelligent technologies and applications. INTAP 2020. Communications in Computer and Information Science. Cham: Springer. https://doi.org/10.1007/978-3-030-71711-7_31
14. Das RK, Islam M, Hasan MM, Razia S, Hassan M, Khushbu SA (2023) Sentiment analysis in multilingual context: comparative analysis of machine learning and hybrid deep learning models. Heliyon 9(9):e20281. https://doi.org/10.1016/j.heliyon.2023.e20281
15. Dang NC, Moreno-García MN, De la Prieta F (2020) Sentiment analysis based on deep learning: a comparative study. Electronics 9:483. https://doi.org/10.3390/electronics9030483
16. Sahoo C, Wankhade M, Singh BK (2023) Sentiment analysis using deep learning techniques: a comprehensive review. Int J Multimed Info Retr 12:41. https://doi.org/10.1007/s13735-023-00308-2
17. Gupta K, Oladimeji D, Varol C, Rasheed A, Shahshidhar N (2023) A comprehensive survey on artifact recovery from social media platforms: approaches and future research directions. Information 14:629. https://doi.org/10.3390/info14120629

18. Uppada SK, Patel P, Sivaselvan B (2022) An image and text-based multimodal model for detecting fake news in OSN's. J Intell Inf Syst. 1–27. https://doi.org/10.1007/s10844-022-00764-y
19. Babu NV, Kanaga EGM (2022) Sentiment analysis in social media data for depression detection using artificial intelligence: a review. SN Comput Sci 3:74. https://doi.org/10.1007/s42979-021-00958-1
20. Sarker IH (2021) Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2:160. https://doi.org/10.1007/s42979-021-00592-x
21. Maliński K, Okarma K (2023) Analysis of image preprocessing and Binarization methods for OCR-based detection and classification of electronic integrated circuit labeling. Electronics 12:2449. https://doi.org/10.3390/electronics12112449
22. MacDermott A, Motylinski M, Iqbal F, Stamp K, Hussain M, Marrington A (2022) Using deep learning to detect social media 'trolls'. Forensic Sci Int: Digit Invest. 43:301446. https://doi.org/10.1016/j.fsidi.2022.301446. ISSN 2666–2817
23. Al-Adhaileh MH, Aldhyani THH, Alghamdi AD (2022) Online troll reviewer detection using deep learning techniques. Appl Bionics Biomech 8(2022):4637594. https://doi.org/10.1155/2022/4637594
24. Michalak H, Okarma K (2019) Improvement of image Binarization methods using image preprocessing with local entropy filtering for alphanumerical character recognition purposes. Entropy (Basel) 21(6):562. https://doi.org/10.3390/e21060562
25. Michalak H, Okarma K (2018) Region based adaptive binarization for optical character recognition purposes. Int Interdiscipl PhD Workshop (IIPhDW) 2018:361–366
26. Yamashita R, Nishio M, Do RKG et al (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9:611–629. https://doi.org/10.1007/s13244-018-0639-9
27. Uzair B, Mehdi M, Sibghat B, Hao T (2023) Editorial: Investigating AI-based smart precision agriculture techniques. Front Plant Sci. 14. https://doi.org/10.3389/fpls.2023.1237783
28. Puttagunta M, Ravi S (2021) Medical image analysis based on deep learning approach. Multimed Tools Appl 80:24365–24398. https://doi.org/10.1007/s11042-021-10707-4
29. 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
30. Anjomshoae S, Omeiza D, Jiang L (2021) Context-based image explanations for deep neural networks. Image Vision Comput. 116:104310. https://doi.org/10.1016/j.imavis.2021.104310. (ISSN 0262-8856)
31. Bhatti U, Mengxing H, Neira-Molin H, Marjan S, Baryalai M, Hao T, Wu G, Bazai S (2023) MFFCG – multi feature fusion for hyperspectral image classification using graph attention network. Expert Syst Appl 229:120496. https://doi.org/10.1016/j.eswa.2023.120496
32. 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(Part A):121282. https://doi.org/10.1016/j.eswa.2023.121282. (ISSN 0957–4174)
33. Valente J, António J, Mora C, Jardim S (2023) Developments in image processing using deep learning and reinforcement learning. J Imaging 9:207. https://doi.org/10.3390/jimaging9100207
34. Nizamani AH, Chen Z, Nizamani AA, Aslam BU (2023) Advance brain tumor segmentation using feature fusion methods with deep U-Net model with CNN for MRI data. J King Saud Univ Comput Inform Sci. 35(9):101793. https://doi.org/10.1016/j.jksuci.2023.101793. (ISSN 1319-1578)
35. Mall PK, Singh PK, Srivastav S, Narayan V, Paprzycki M, Jaworska J, Ganzha M (2023) A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. Healthcare Analyt. 4:100216. https://doi.org/10.1016/j.health.2023.100216. (ISSN 2772-4425)
36. Li X, Cui M, Li J, Bai R, Lu Z, Aickelin U (2021) A hybrid medical text classification framework: Integrating attentive rule construction and neural network. Neurocomputing. 443:345–355. https://doi.org/10.1016/j.neucom.2021.02.069. (ISSN 0925-2312)
37. Naithani K, Raiwani YP (2023) Realization of natural language processing and machine learning approaches for text-based sentiment analysis. Expert Syst 40(5):e13114. https://doi.org/10.1111/exsy.13114
38. JayaLakshmi ANM, Kishore KV (2022) Performance evaluation of DNN with other machine learning techniques in a cluster using Apache Spark and MLlib. J King Saud Univ- Comput Inform Sci. 34(1):1311–1319. https://doi.org/10.1016/j.jksuci.2018.09.022. (ISSN 1319-1578)
39. Yenkikar A, Babu CN, Hemanth DJ (2022) Semantic relational machine learning model for sentiment analysis using cascade feature selection and heterogeneous classifier ensemble. PeerJ Comput Sci 20(8):e1100. https://doi.org/10.7717/peerj-cs.1100
40. Elahi M, Afolaranmi SO, Martinez Lastra JL et al (2023) A comprehensive literature review of the applications of AI techniques through the lifecycle of industrial equipment. Discov Artif Intell 3:43. https://doi.org/10.1007/s44163-023-00089-x
41. Androcec D (2020) Machine learning methods for toxic comment classification: a systematic review. Acta Universitatis Sapientiae, Informatica 12:205–216. https://doi.org/10.2478/ausi-2020-0012
42. Rahul, Kajla H, Jatin H, Gajanand S (2020) Classification of online toxic comments using machine learning algorithms. 1119–1123. https://doi.org/10.1109/ICICCS48265.2020.9120939.
43. Čepulionytė A, Toldinas J, Lozinskis B (2023) A multilayered preprocessing approach for recognition and classification of malicious social network messages. Electronics 12:3785. https://doi.org/10.3390/electronics12183785
44. Belfield SJ, Cronin MTD, Enoch SJ, Firman JW (2023) Guidance for good practice in the application of machine learning in development of toxicological quantitative structure-activity relationships (QSARs). PLoS ONE 18(5):e0282924. https://doi.org/10.1371/journal.pone.0282924
45. Abbasi A, Javed AR, Iqbal F, Kryvinska N, Jalil Z (2022) Deep learning for religious and continent-based toxic content detection and classification. Sci Rep 12(1):17478. https://doi.org/10.1038/s41598-022-22523-3
46. Asudani DS, Nagwani NK, Singh P (2023) Impact of word embedding models on text analytics in deep learning environment: a review. Artif Intell Rev 56:10345–10425. https://doi.org/10.1007/s10462-023-10419-1
47. Danilo D, Recupero R, Diego, Harald S (2021) An assessment of deep learning models and word embeddings for toxicity detection within online textual comments. Electronics. 10. https://doi.org/10.3390/electronics10070779.
48. Ashok Kumar J, Abirami S, Trueman TE, Cambria E (2021) Comment toxicity detection via a multichannel convolutional bidirectional gated recurrent unit. Neurocomputing. 441:272–278. https://doi.org/10.1016/j.neucom.2021.02.023. (ISSN 0925-2312)
49. Maslej-Krešňáková V, Sarnovský M, Butka P, Machová K (2020) Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification. Appl Sci 10:8631. https://doi.org/10.3390/app10238631
50. Jahan MdS, Oussalah M (2023) A systematic review of hate speech automatic detection using natural language processing. Neurocomputing. 546:126232. https://doi.org/10.1016/j.neucom.2023.126232. (ISSN 0925-2312)
51. Mehendale N, Shah K, Phadtare C, Rajpara K. Cyber bullying detection for Hindi-English language using machine learning (May 21, 2022). Available at SSRN: https://ssrn.com/abstract=4116143 Or https://doi.org/10.2139/ssrn.4116143
52. Alruily M (2021) Classification of Arabic tweets: a review. Electronics 10:1143. https://doi.org/10.3390/electronics10101143

Publisher's Note