RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 4, OCTOBER - DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

AN AI – DRIVEN FRAMEWORK FOR DETECTING EXPLICIT AND IMPLICIT ABUSIVE COMMENTS IN ONLINE SOCIAL MEDIA

Dr. M. Regina [1], Hamroudhavarshiny R[2]
[1] Assistant Professor, Loyola College, Chennai–600034, Tamil Nadu, India
[2] M.Sc. Student, Loyola College, Chennai–600034, Tamil Nadu, India
Email: reginamathew@loyolacollege.edu; hamroudhavarshiny@gmail.com;

ABSTRACT

The rapid growth of online social media has led to a significant increase in abusive and harmful interactions, making automated content moderation in important research challenge. Although existing machine learning and deep learning techniques have achieved promising results in detecting explicit abusive comments, they often struggle to identify implicit abusive comments expressed through subtle, indirect or context – dependent language. This limitation reduces the effectiveness of current moderation systems and allows harmful content to remain undetected. The proposed research aims to develop an AI – Driven framework for detecting both explicit and implicit abusive comments by comparatively evaluating multiple machine learning and deep learning models. The research will involve collecting benchmark datasets, preprocessing textual data, training and evaluating different models using standard performance metrics, and identifying the most effective approach for abusive comment detection. The selected model will be further validated using real world online comments to assess its practical applicability in intelligent content moderation. The proposed framework is expected to improve the detection of subtle abusive language, support safer online interactions, and provide valuable insights for enhancing automated moderations systems across digital platforms.

Keywords*: Explicit Abuse, Implicit Abuse, Machine Learning, Deep Learning, Natural Language Processing, Text Classification, Social Media*

1. INTRODUCTION

Online social media platforms have become an integral part of modern communication, allowing users to exchange opinions, engage with others, and participate in public debates. However, the proliferation of these platforms has also contributed to an increase in abusive and harmful interactions. Hateful comments can negatively impact individuals and communities online, creating the need for powerful automated approaches to identify and

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

contain harmful speech. There is more than one form of swearing. Explicitly offensive comments typically include direct insults, offensive language, profanity, or overtly offensive language. In contrast, implicitly offensive comments may convey harmful or offensive meanings indirectly through contextual, subtle, or implicit language. Direct insults are often identified by recognizable language patterns, whereas indirect insults are more difficult to detect because the offensive meaning may not be expressed explicitly in the sentence. This distinction makes the classification of overt and covert abuse an important issue in automated content moderation.

Machine Learning and Deep Learning techniques are widely used for text classification and misuse detection because they can learn linguistic patterns from text data. While traditional machine learning approaches can represent textual information using statistical features such as Term Frequency or Inverse Document Frequency. Deep Learning approaches can learn sequential and contextual patterns from textual representations. The performance of these approaches may vary depending on dataset characteristics, preprocessing methods, feature representation, model architecture, and target class distribution. Existing approaches to detecting profanity often focus on determining whether a comment is offensive, and distinguishing between different forms of insults has received relatively little attention. In particular, determining whether offensive comments are overt or covert poses further classification problems. A system that only considers the presence of offensive words may not recognize indirect offensive language well, whereas a model that can learn a broader range of text patterns may provide additional information to distinguish between the two categories.

This study focuses on the classification of offensive comments into Explicit and Implicit. A dataset of manually labelled offensive comments was processed and used to develop and evaluate several Machine Learning and Deep Learning models. Machine Learning approaches include Logistic Regression, Linear Support Vector Machine, Random Forest, XGBoost and tuned XGBoost. Deep Learning approaches include Long Short – Term Memory (LSTM), Bidirectional Long Short – Term Memory, and Gated Recurrent Unit. Machine Learning models were trained using TF – IDF based text representation, and deep learning models used tokenized and padded text sequences. We compared the models using standard classification metrics such as Precision, Accuracy, Recall and F1 – Score. This evaluation not only provides a basis for testing the effectiveness of various machine learning and deep learning approaches in distinguishing implicit and explicit offensive language, but also allows us to identify

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

challenges associated with the classification of Implicit language. The rest of this document is organized as follows: Section 2 presents Related work on profanity detection and classification of Implicit and Explicit abuse. Section 3 describes the methodology, including datasets, preprocessing, feature representation, model development, and evaluation procedures. Section 4 presents a comparison of the experimental results and the performance of the evaluation model, discusses the findings and limitations of the study. Finally, Section 5 concludes the study and provides directions for future work.

2. RELATED WORK

2.1 Existing Research

Abusive language detection has been widely studied using Machine Learning (ML) and Deep Learning (DL) techniques. Early approaches primarily focused on detecting offensive or abusive content using traditional classifiers such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and ensemble methods with features such as TF – IDF and Bag – of – Words. Alam et al. (2021) demonstrated that ensemble ML methods can improve offensive – language classification, while Talpur and O'Sullivan (2020) showed that feature engineering and class – balancing techniques can improve cyberbullying detection.

Recent research has given greater attention to implicit hate and abuse, where harmful meaning is expressed indirectly. ElSherief et al. (2021) introduced the Latent Hatred benchmark and showed that implicit hate can involve sarcasm, coded language, and metaphor, making it more difficult to classify than explicit hate. Lin (2022) further examined the use of external world knowledge for implicit hate detection and found that such knowledge does not consistently improve the recognition of nuanced implicit expressions.

Deep Learning and Transformer – Based approaches have also been increasingly explored. Prama et al. (2025) evaluated LSTM – based cyberbullying severity detection, while Adhikari et al. (2026) reported stronger toxic – comment classification performance using fine – tuned BERT compared with traditional ML models. Kim et al. (2024) investigated contrastive learning with hard – negative sampling and demonstrated the importance of distinguishing semantically similar but differently labelled examples.

Several studies have also identified limitations affecting abusive – language detection. Yin and Zubiaga (2021) highlighted problems such as dataset bias, limited generalization, non –

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

standard language, and implicit expressions. Zhang et al. (2024) showed that even LLMs can struggle with implicit hate and may produced poorly calibrated predictions. These findings indicate that, despite advances in Machine Learning and Deep Learning, distinguishing explicit and implicit abusive language remains a challenging classification problem.

2.2 Research Gap

The review of the existing studies highlights a number of research gaps related to the detection of abusive language. The studies are dominated by solutions that attempt to identify abusive and non – abusive texts. Yet, there appears to be a lack of research into the detection of explicit and implicit abusive content. The latter category is particularly challenging because it often relies on implicit meaning that conveys abusive content via sarcasm, hidden insults, and other implicit formulations.

The review also suggests that traditional Machine Learning methods offer little help in addressing the challenge of abusive language detection. While TF – IDF representations are accurate and computationally inexpensive, they are not conducive to capturing complex interrelationships in natural language. More advanced solutions, such as Deep Learning and related approaches, appear to be more promising, yet they require substantial computational power and encounter a number of challenges identified by researchers.

Finally, the review highlights the lack of comparative analysis between traditional Machine and Deep Learning solutions. Indeed, there are few studies that attempt to compare the effectiveness of various models in Binary classification tasks of implicit and explicit abusive content detection.

3. METHODOLOGY

The methodology adopted in this study involves a pipeline of classifying abusive comments into Explicit Abuse and Implicit Abuse. The experimental process was done through preparing a dataset, text pre – processing, splitting of data into testing and training sets, feature and text representation, developing a Machine Learning (ML) and Deep Learning (DL) model, training the model and evaluating its performance. The same data was used in all the processes to allow the comparison of classification performance of ML and DL approaches.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

3.1 Dataset Preprocessing

The dataset used in this study consists of 1,141 manually labelled abusive comments. The two target categories are Explicit Abuse and Implicit Abuse. Explicit Abuse refers to comments in which abusive meaning is explicitly stated; Implicit Abuse refers to comments in which abusive meaning is implicitly suggested or requires more contextual information and interpretation to distinguish from non – abusive comments.

692 comments were labeled as Explicit Abuse and 449 comments were labeled as Implicit Abuse.

The comment text underwent preprocessing before being used to train the model. A cleaned version of each comment was generated and saved in the 'clean_comment' field. The text was converted to lowercase to ensure consistency between variations of the same word with different capitalization. Irrelevant variations of the same comment were removed to retain the relevant linguistic information contained in the comment that would be useful in differentiating between the two types of abuse.

The preprocessed dataset was then divided into training, validation, and test sets in a 70:15:15 ratio. This resulted in 798 training samples, 171 validation samples, and 172 test samples. The training set was used to train the model, while the validation set was used to compare different models. The test was used to evaluate the final performance of the chosen model after training.

3.2 Feature Extraction

Various text – encoding techniques were applied based on the requirements of respective ML and DL algorithms. In particular, to fit into an ML paradigm, comments were converted into numerical TF – IDF representation. A TF – IDF (Term Frequency – Inverse Document Frequency) encoding technique quantifies the importance of certain terms in a comment relative to their overall frequency in a dataset. The vectorizer was initialized with a maximum of 5000 features to control the dimensional complexity of text representation. By using TF – IDF encoding, a fixed – length numerical vectorization of comments was achieved, suitable for conventional ML algorithms.

In turn, numerical representations of comments in DL were constructed from integer – encoded word sequences. First, a Tokenizer was fitted on the training set with a maximum vocabulary size of 5000 words. During the encoding process, rare words not present in the tokenizer's

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

vocabulary were assigned a special out – of – vocabulary token. Since comments included different number of words, sequences were then padded or truncated to a fixed length of 100 tokens.

3.3 Machine Learning Models Development

The initial pool of ML algorithms included in the consideration consisted of five classifiers: Logistic Regression, Linear Support Vector Machine (SVM), Random Forest, XGBoost, and tuned XGBoost. All algorithms were applied to the TF – IDF matrix.

The first two algorithms were the linear classifiers applied to detect patterns in the input data. The Random Forest algorithm was chosen as the representative of ht tree – based approach, while XGBoost was used as a gradient – boosting alternative that is reported to provide better modelling power than tree – based algorithms on average. The last algorithm that was applied, which also required the largest computational expenses, was the tuned XGBoost. It was used to optimize the following parameters: number of estimators, max tree depth, learning rate, subsampling ratio, and feature – subsampling ratio. Overall, it was used to improve the performance of the classifier compared to the default settings.

3.4 Deep Learning Models Development

Three Deep Learning models: Long Short – Term Memory, Bidirectional Long Short – Term Memory and Gated Recurrent Unit was explored. All the networks were taking as input the tokenized and padded comment sequences and were trained to classify the texts based on their learned sequential patterns.

The Long Short – Term Memory network was employing an embedding layer to encode the word indices as dense vectors followed by a LSTM layer and fully connected layers to perform the classification. The Bidirectional Long Short – Term Memory networks processed the input sequence in both directions, thus enabling the network to learn patterns from both directions of the input. Finally, Gated Recurrent Unit was used to learn sequential patterns of the comment’s texts.

The dropout regularization was applied in order to prevent the Deep – Learning models from overfitting on the training samples. Another technique to address the overfitting issue was to balance the class weights in the final deep learning model since the Explicit Abuse class was much more represented than the Implicit Abuse class.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

3.5 Model Training and Performance Testing

The ML and DL models were trained on the training subset and validated using the validation subset. The latter was used to evaluate the models against each other and, in the case of XGBoost, to determine the impact of hyperparameter optimization. The performance of the models was finally tested using the independent test subset

The performance of the models was evaluated using Accuracy, Precision, Recall, and F1 – score. Accuracy represents the ratio of correctly classified comments to all comments. In turn, Precision is the ratio of correctly positively classified comments to all positively classified comments. Finally, Recall is the ratio of correctly positively classified comments to all actual positive comments, and the F1 – score is the harmonic mean of Precision and Recall. Additionally, confusion matrices were analyzed to highlight the difference between Explicit Abuse and Implicit Abuse

The same performance criteria were used to compare ML and DL models, enabling a consistent evaluation of their classification performance.

4. RESULTS & DISCUSSION

The performance of the Machine Learning (ML) and Deep Learning (DL) models was evaluated using the independent test dataset containing 172 comments. Accuracy, Precision, Recall, and F1 – Score were used as the primary evaluation measures. Confusion matrices were additionally examined to analyze the classification behavior of the models, particularly their ability to distinguish between Explicit Abuse and Implicit Abuse.

4.1 Machine Learning Results

The test performance of the five Machine Learning models is presented in Table 1. All ML models were evaluated using the same TF – IDF based test representation.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 – Score (%)**
Logistic Regression	70.93	62.86	64.71	63.77
Linear Support Vector Machine	70.93	65.00	57.35	60.94
Random Forest	66.28	61.36	39.71	48.21
XGBoost	70.93	63.24	63.24	63.24
Tuned XGBoost	71.51	64.62	61.76	63.16

Table 1. Test Performance of Machine Learning Models

RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 4, OCTOBER - DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

The best – performing model, XGBoost – tuned, achieved the highest test accuracy among the tested ML models, scoring 71.51%. In turn, the Logistic Regression, Linear SVM, and XGBoost algorithms yielded 70.93%, while the Random Forest method scored 66.28%.

The confusion matrices below demonstrate the number of correctly predicted Implicit and Explicit comments for each of the tested algorithms. Thus, the Logistic Regression algorithm correctly classified 78 Explicit and 44 Implicit comments. The Linear SVM classifier achieved 83 Explicit and 39 Implicit comments. In turn, the Random Forest method allowed classifying 87 Explicit and only 27 Implicit comments. Meanwhile, the best – performing XGBoost – tuned classifier achieved 81 Explicit and 42 Implicit comments.

The confusion matrix for the best – performing XGBoost – tuned algorithm is demonstrated below in Figure 1. This algorithm correctly predicted 81 Explicit and 42 Implicit comments, while classifying 23 Explicit comments as Implicit and 26 Implicit comments as Explicit.

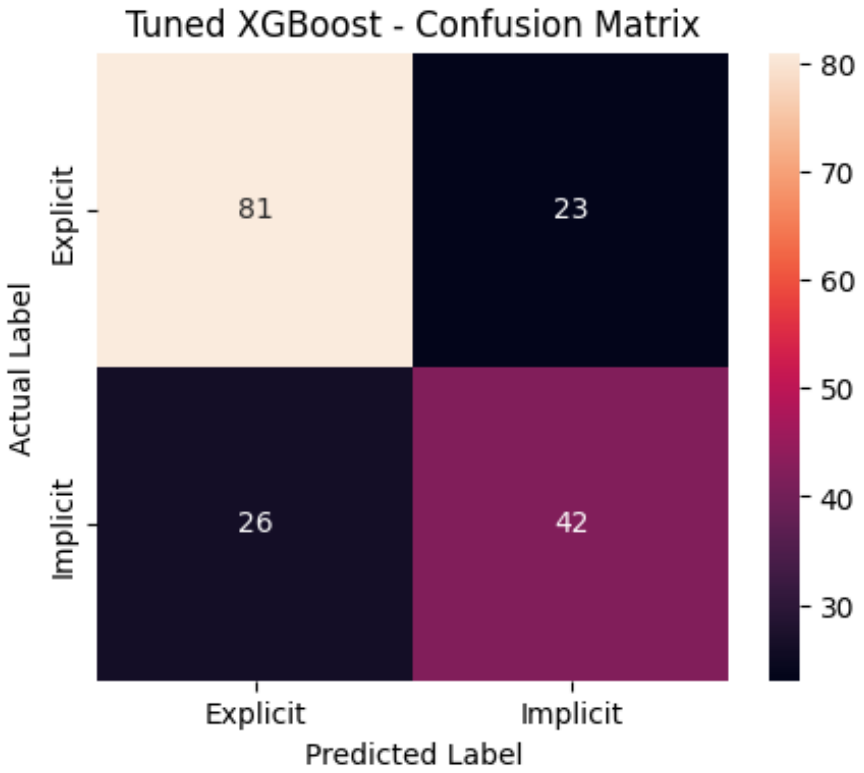


Figure 1. Confusion matrix of the tuned XGBoost model

RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 4, OCTOBER - DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

4.2 Deep Learning Results

The Deep Learning models were evaluated using tokenized and padded comment sequences. The resulting test performance is presented in Table 2.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 – Score (%)
LSTM	60.47	36.56	60.47	45.57
BiLSTM	70.93	70.68	70.93	70.76
GRU	60.47	36.56	60.47	45.57

Table 2. Test performance of Deep Learning Models

Among the evaluated Deep Learning models, BiLSTM achieved the highest accuracy of 70.93%, together with a weighted F1 – score of 70.76%. LSTM and GRU both achieved an accuracy of 60.47%

The confusion matrices reveal an important difference I the behavior of the recurrent models. The LSTM and GRU models classified all 104 Explicit comments correctly but classified all 68 Implicit comments as Explicit. Consequently, these models did not identify any Implicit comments in the test set. In contrast, the BiLSTM model correctly classified 81 Explicit and 41 Implicit comments, while misclassifying 23 Explicit and 27 Implicit comments.

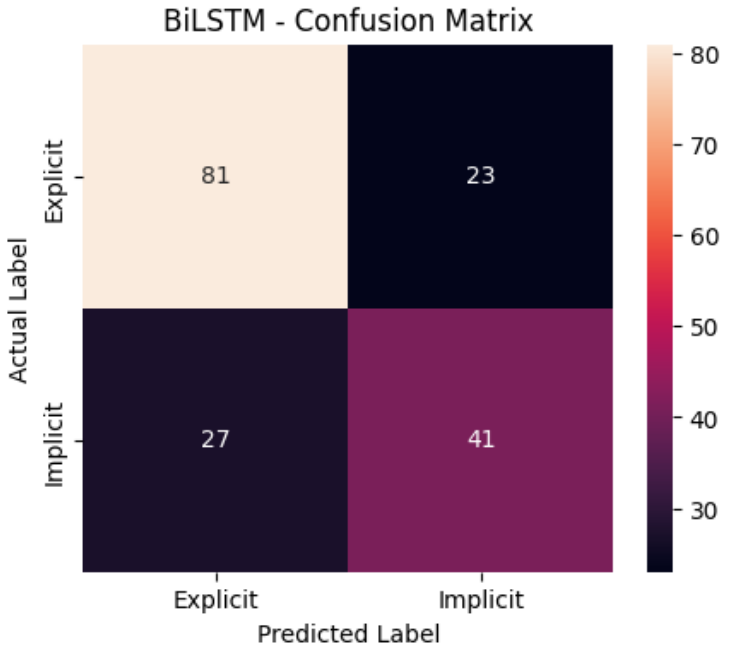


Figure 2. Confusion matrix of the BiLSTM model

RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 4, OCTOBER - DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

4.3 Comparative Analysis of ML & DL Models

The overall test accuracy of the evaluated ML and DL models is presented in Table 3.

Category	Model	Test Accuracy (%)
ML	Logistic Regression	70.93
ML	Linear SVM	70.93
ML	Random Forest	66.28
ML	XGBoost	70.93
ML	Tuned XGBoost	71.51
DL	LSTM	60.47
DL	BiLSTM	70.93
DL	GRU	60.47

Table 3. Comparative test accuracy of all evaluated models

The comparative results indicate that the evaluated models produced different levels of performance on the test dataset. The tuned XGBoost models recorded the highest test accuracy of 71.51%, while BiLSTM achieved 70.93% and provided the strongest performance among the evaluated Deep Learning models. Logistic Regression, Linear SVM, and XGBoost also achieved 70.93%

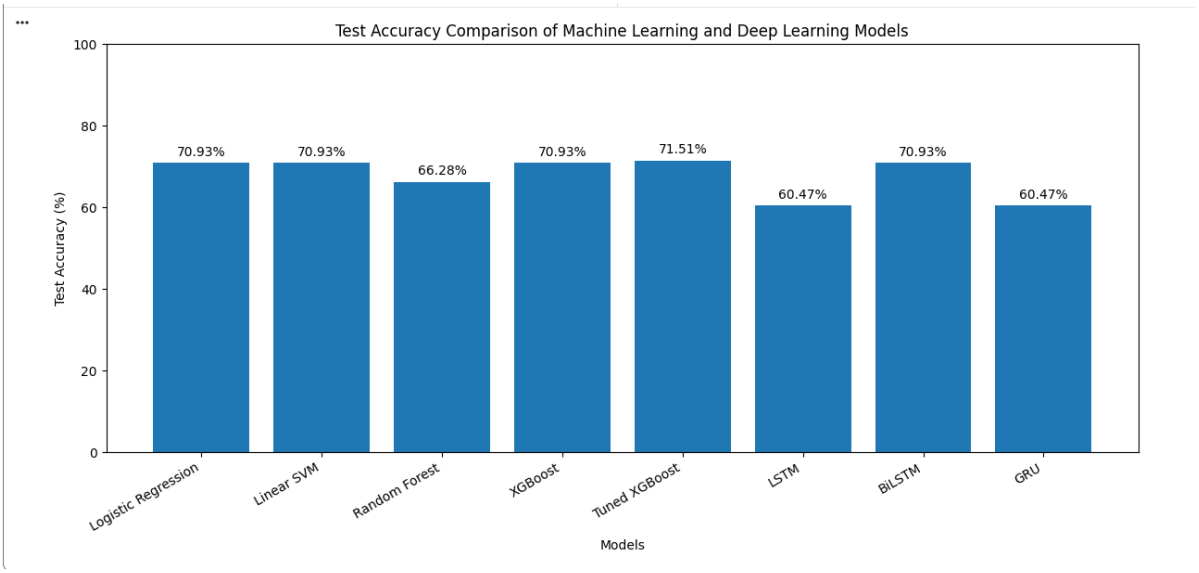


Figure 3. Test accuracy comparison of Machine Learning and Deep Learning models.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

The results also demonstrate that accuracy alone does not fully describe the behavior of the models for this classification task. In particular, the LSTM and GRU models achieved 60.47% accuracy while failing to identify any Implicit Abuse comments. In contrast, the BiLSTM model identified 41 of the 68 Implicit Abuse comments in the test set. This highlights the importance of examining class – level performance and confusion matrices alongside overall accuracy when evaluating explicit and implicit abuse classification.

4.4 Performance of Machine Learning and Deep Learning Models

The results of the analysis prove that both classical Machine Learning and Deep Learning approaches can be successfully applied in the solution of the classification tasks, including the case of explicit and implicit abusive language. However, their accuracy depends significantly on the choice of the proper method for representation learning and the volume of the training set.

Machine Learning methods used in the research employed the TF – IDF technique, which allows highlighting the words and terms characteristic of each category, thus, adequately representing this type of information. This approach is relatively simple; however, it's performance is conditional to the presence of statistical regularities in the input data. On the other hand, Deep Learning methods process sequences of tokens and automatically detect patterns specific to each category with the help of a series of transformations performed by neural networks. It makes them highly accurate, but only with sufficient training data and proper model selection.

Thus, all the conducted research can lead to the conclusion that there is no direct connection between the complexity of the model and its accuracy in the context of text classification tasks. A more complex model will not necessarily perform better if the training data is insufficient or the target class distinction is not encoded in the data representation method. Therefore, when selecting a model for a specific task, in addition to the volume and nature of the data, the choice of representation method and the properties of the classes should be considered.

4.5 Limitations

The study has several limitations, that should be considered when interpreting the results. The dataset used for manual labelling consists of only 1,141 comments. This represents a relatively small amount of labelled data for training complex Deep Learning architectures. The

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

distribution of the two classes is also not completely balanced, with Explicit abuse comments being more numerous that Implicit abuse comments.

Another limitation is that the classification of implicit abuse can be subjective. The meaning of an indirection can sometimes not be apparent from the text, leading to ambiguity in manual labelling and classification by models.

Furthermore, the experiments were conducted using a single dataset and a particular preprocessing and modelling configuration. As such, it is not possible to extrapolate the observed performance to comments on different social – media platforms, communities, languages, or domain. A more extensive evaluation on additional datasets and an external validation would have been required to investigate the approaches generalization further.

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

This study developed an AI Framework for classifying abusive comments into Explicit Abuse and Implicit Abuse. A manually labelled set of abusive comments was prepared and processed, and both Machine Learning and Deep Learning approaches were applied to classify abuse.

The developed framework used TF – IDF based feature representation for traditional Machine Learning and tokenized sequences for the Deep Learning approach. Multiple classification approaches were developed and evaluated using comparable evaluation measures. The experimental results show that automated classification of explicit and implicit abuse is possible but challenging, and that abusive meaning that is expressed indirectly or contextually, is difficult.

Overall, this study demonstrated that combining different approaches of Machine Learning and Deep Learning can allow for fine – grained abusive – language analysis. Instead of considering abusive content as on category, differentiating between explicit abuse and implicit abuse, provides a more informative view for automated content moderation. The results can provide a foundation for developing more context – aware automated approaches for detecting different forms of abusive language in online social – media environments.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

5.2 Future Work

There are several directions this study can go. A larger and more diverse set of comments, from multiple social – media platforms can be developed to increase the generalization of the framework. More manually labelled examples of implicit abuse would also enable training data that allows for learning subtle and contextually dependent patterns.

Future studies can consider Transformer – Based architectures and other language models that can capture rich semantic relations. Conversational information, replies, surrounding comments, and other contextual information can also aid identifying abusive intent that can only be inferred from an individual comment. Further research can also examine improving handling of sarcasm, indirect expressions, coded language, and ambiguous comments.

Evaluations using external datasets and social – media data would also assess how well the developed approaches would generalize beyond the dataset used. These improvements would result in more robust and context – aware automated moderation approaches for detecting different forms of abusive language and improving moderation for online communities.

References

[1.] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury and D. Yang, “Latent Hatred: A Benchmark for Understanding Implicit Hate Speech,” *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 345–363. doi: 10.18653/v1/2021.emnlp-main.29.

[2.] B. A. Talpur and D. O’Sullivan, “Cyberbullying Severity Detection: A Machine Learning Approach,” *PLOS ONE*, vol. 15, no. 10, 2020, e0240924. doi: 10.1371/journal.pone.0240924.

[3.] T. T. Prama, J. F. Amrin, M. M. Anwar and I. H. Sarker, “AI-Enabled User-Specific Cyberbullying Severity Detection with Explainability,” *arXiv preprint arXiv:2503.10650*, 2025.

[4.] K. S. Alam, S. Bhowmik and P. R. K. Prosun, “Cyberbullying Detection: An Ensemble Based Machine Learning Approach,” *Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 2021. doi: 10.1109/ICICV50876.2021.9388499.

[5.] J. Kim, S. Jin, S. Park, S. Park and K. Han, “Label-aware Hard Negative Sampling Strategies with Momentum Contrastive Learning for Implicit Hate Speech Detection,” *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 16177–16188. doi: 10.18653/v1/2024.findings-acl.957.

[6.] W. Yin and A. Zubiaga, “Towards Generalisable Hate Speech Detection: A Review on Obstacles and Solutions,” *PeerJ Computer Science*, vol. 7, e598, 2021. doi:

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/007

10.7717/peerj-cs.598.

[7.] M. Zhang, J. He, T. Ji and C.-T. Lu, "Don't Go To Extremes: Revealing the Excessive Sensitivity and Calibration Limitations of LLMs in Implicit Hate Speech Detection," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024, pp. 12073–12086. doi: 10.18653/v1/2024.acl-long.652.

[8.] J. Lin, "Leveraging World Knowledge in Implicit Hate Speech Detection," *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, 2022, pp. 31–39. doi: 10.18653/v1/2022.nlp4pi-1.4.

[9.] B. Adhikari, P. Sapkota and R. Shakya, "Multi-Label Toxic Comment Detection Using BERT-Based NLP and Machine Learning," *International Journal on Engineering Technology and Infrastructure Development (InJET-InDev)*, vol. 2, no. 2, pp. 37–45, 2026. doi: 10.3126/injet-indev.v2i2.95699.

[10.] C. Duchêne, H. Jamet, P. Guillaume and R. Dehak, "A Benchmark for Toxic Comment Classification on Civil Comments Dataset," *arXiv preprint arXiv:2301.11125*, 2023. doi: 10.48550/arXiv.2301.11125.

[11.] R. Alex, M. Subramanian, J. J. R, K. B. A. T. and T. Mukherjee, "Learning without Forgetting in NLP: Deep Learning Based Transformer Models for Lifelong Task Incremental Learning," *Edelweiss Applied Science and Technology*, vol. 9, no. 11, pp. 448–466, 2025. doi: 10.55214/2576-8484.v9i11.10905.

[12.] N. B. Ocampo, E. Sviridova, E. Cabrio and S. Villata, "An In-depth Analysis of Implicit and Subtle Hate Speech Messages," *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 1997–2013. doi: 10.18653/v1/2023.eacl-main.147.

[13.] S. Ghosh, M. Suri, P. Chiniya, U. Tyagi, S. Kumar and D. Manocha, "CoSyn: Detecting Implicit Hate Speech in Online Conversations Using a Context Synergized Hyperbolic Network," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 6159–6173. doi: 10.18653/v1/2023.emnlp-main.377.

[14.] S. Roychowdhury and V. Gupta, "Data-Efficient Methods For Improving Hate Speech Detection," *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 125–132. doi: 10.18653/v1/2023.findings-eacl.9.

[15.] T. Bourgeade, Z. Li, F. Benamara, V. Moriceau, J. Su and A. Sun, "Humans Need Context, What about Machines? Investigating Conversational Context in Abusive Language Detection," *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 8438–8452.

[16.] H. Kirk, B. Vidgen and S. A. Hale, "Is More Data Better? Re-thinking the Importance of Efficiency in Abusive Language Detection with Transformers-Based Active Learning," *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, 2022, pp. 52–61.