

Toxic Comment Classification

A review paper

****

**AUTHORS:**TRISHA SHARMA 229309215

ANAHITA BHANDARI 229309186

**DEPARTMENT OF CSE**

**SCHOOL OF INFORMATION, SECURITY AND DATA SCIENCE**

November 15, 2024

Manipal University Jaipur

**TOXIC COMMENT CLASSIFICATION- A REVIEW PAPER**

***Abstract*-** *The increase in social media usage and online platforms has led to a notable rise in user-created content, improving communication but also giving rise to negative actions such as hate speech, harassment, and inappropriate comments. Such negative comments have the potential to severely harm online conversations and the overall well-being of the community. This research presents a machine learning technique for automatically categorizing toxic comments, aiming to develop an effective tool for identifying and organizing harmful content in large datasets. We explore various machine learning techniques, such as logistic regression, support vector machines, and neural networks, by analysing a diverse collection of labelled comments. The results suggest that these models can enhance online moderation efforts and promote more positive interactions in digital spaces. This research contributes to the ongoing discussion about improving online safety through automated systems, addressing challenges with language complexity, data imbalance, and contextual sensitivity.*

1. **INTRODUCTION**

In the digital era, internet platforms act as strong tools for communication, collaboration, and self-expression. Nevertheless, with the numerous chances for engagement and discussion, the rise of harmful behaviour in online communities has become a significant issue. Harmful remarks, including hate speech, harassment, insults, and various types of detrimental content, diminish the quality of online interactions and present serious threats to personal well-being and community unity. Consequently, the categorization of harmful remarks has attracted considerable interest from scholars, platform overseers, and decision-makers as well.

The primary objective of toxic comment classification is to develop computational models capable of automatically distinguishing between toxic and non-toxic comments in vast quantities of user-created content by effectively recognizing harmful behaviour, these models assist platform moderators in efficiently overseeing conversations, safeguarding users from abuse, and fostering a constructive online atmosphere. Additionally, the classification of toxic comments supports wider societal objectives, such as encouraging free speech, curbing cyberbullying, and maintaining online civility. By systematically analysing and categorizing toxic comments, researchers can understand the frequency, dynamics, and root causes of toxic behaviour, guiding interventions and policy choices that promote safer and more inclusive online environments.

This review article focuses on the difficulties in classifying toxic remarks because of the subtle complexities of language and the varied forms of toxicity. Major challenges involve linguistic intricacies, as toxic comments often feature delicate cues like sarcasm and cultural subtleties; data imbalance, with harmful remarks usually making up a lesser segment of datasets compared to non-toxic ones; context sensitivity, as interpretations of toxicity differ among communities; and model interpretability, crucial for guaranteeing transparency in automated moderation choices.

The study explores current machine learning methods for classifying toxic comments, evaluating different models and natural language processing strategies utilized in this domain. Through examining research that employs various datasets of labelled comments and sophisticated NLP techniques, we intend to emphasize the efficacy of existing approaches along with their drawbacks. This review offers important perspectives on the intricacies of online toxicity, seeking to guide future studies in creating more dependable and transparent online moderation systems while promoting healthier digital interactions.

1. **LITERATURE REVIEW**

Identifying and reducing harmful online interactions is a crucial research focus in the field of toxic comment classification. This review of literature talks about the difficulties linked to identifying toxic comments, current methods using diverse machine learning and natural language processing methods, strategies for managing data, and the significance of understanding model decisions. Furthermore, it emphasizes frequently utilized datasets in this field.

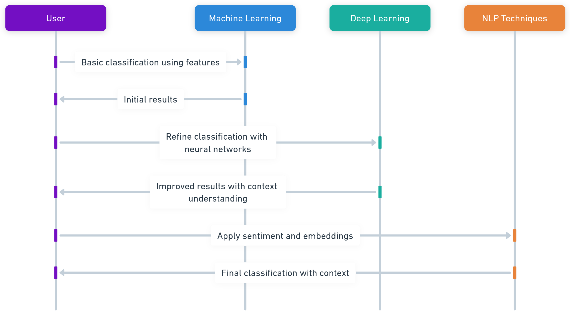
2.1 Challenges in Toxic Comment **Classification:**

1. Linguistic Complexity: Toxic comments often exhibit diverse linguistic patterns that include sarcasm, irony, and cultural nuances. This complexity makes it challenging for models to accurately capture subtle cues indicative of toxicity. The variability in language use across different demographics and contexts further complicates the classification process.
2. Data Imbalance: A significant challenge in toxic comment classification is the imbalance between toxic and non-toxic comments in available datasets. Toxic comments are typically much rarer than their non-toxic counterparts, leading to biased model performance. This imbalance can result in models that are overly conservative, failing to identify toxic comments effectively.
3. Context Sensitivity: The interpretation of toxicity can vary widely across different contexts, communities, and cultural norms. A comment deemed toxic in one setting may be acceptable in another, necessitating adaptable classification strategies that consider contextual nuances and user perceptions.
4. Model Interpretability: As automated moderation systems increasingly influence user experiences, ensuring transparency and interpretability of classification models is vital. Users need to trust these systems, which requires clear explanations of how decisions are made regarding comment classifications.
   1. **Existing Approaches**

**Machine learning:** Classic machine learning techniques such as Support Vector Machines (SVM) and decision trees have been widely applied for categorizing toxic comments**.** These models generally depend on feature extraction techniques to convert textual data into numerical representations. Although useful for preliminary investigations, these models might face challenges with the complexities of language without significant feature engineering.

**Deep Learning Methods:** Deep learning techniques are popular due to their ability to grasp complex patterns from large amounts of data. Neural networks, particularly Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), have shown great success in categorizing offensive remarks by grasping the sequential connections and context in the text. Additionally, BERT and GPT, which are transfer learning models, have revolutionized the field by providing pre-trained embeddings that capture a wide range of linguistic features, significantly improving classification accuracy.

**Natural Language Processing (NLP) Techniques:** The utilization of NLP techniques is crucial for categorizing toxic remarks. Approaches such as sentiment analysis help in evaluating the emotional aspect of comments, while word embeddings like Word2Vec and GloVe provide vector representations of words that indicate semantic relationships. Transformers have become a robust framework for handling text data, allowing models to comprehend context more effectively than conventional approaches.



***Fig.1 Sequence Diagram of existing approaches***

**Data Handling Strategies:**

To address data imbalance challenges, several strategies have been implemented:

* **Oversampling**: This method entails creating copies of the minority class (toxic comments) to even out the dataset.
* **Under sampling**: On the other hand, under sampling involves decreasing the quantity of cases in the majority group (non-toxic comments) in order to create equilibrium.
* **Synthetic Data Generation**: Methods like SMOTE generate artificial instances of underrepresented categories to augment training data by avoiding mere replication of current samples.

These strategies aim to improve model performance by providing a more balanced representation of classes during training.

**2.3 Interpretability and Explainability**

As machine learning models become integral to moderating online content, enhancing their interpretability is essential for fostering user trust. Various methods have been developed to improve transparency in model decisions:

* LIME (Local Interpretable Model-agnostic Explanations): This technique provides local explanations for individual predictions by approximating the model's behaviour around a specific instance.
* SHAP (SHapley Additive exPlanations): SHAP values offer a unified measure of feature importance based on cooperative game theory principles, helping users understand which features influenced a model's decision.

**2.4 Datasets for Toxic Comment Classification**

Several datasets are commonly used for training and evaluating toxic comment classification models:

* **Jigsaw Dataset**: This dataset includes comments from online platforms labelled with various toxicity levels, making it a widely utilized resource for training classifiers.
* **Wikipedia Comments Dataset**: This dataset comprises user comments from Wikipedia discussions, annotated for different types of toxicity.
* **Kaggle Toxic Comment Classification Challenge Dataset**: A popular dataset that contains comments labelled as toxic or non-toxic across multiple categories such as hate speech and threats.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author Name and Reference No.** | **Major Findings** | **Advantages** | **Major Shortcomings** | **Datasets/Tools or Software Used** |
| Chaitanya Sonawane et al., | Created a system based on deep learning for categorizing toxic comments utilizing SVM and NLP methods. | Improves online safety by swiftly detecting harmful content, enhancing user experience. | Restricted to the datasets employed; might not apply effectively across various platforms. | Deep learning frameworks, SVM, NLP techniques (e.g., TF-IDF, Word2Vec) |
| Bala Durga Dharmavarapu & Jayanag Bayana | Proposed a sarcasm detection system using Naive Bayes and AdaBoost for Twitter data. | Improves sentiment analysis accuracy by distinguishing sarcastic comments from non-sarcastic ones. | Depends significantly on feature selection; could overlook subtleties in sarcasm. | Twitter dataset, Naive Bayes, AdaBoost |
| Zeel Doshi,subhash Nadkarni,Kushal Ajmera,Neepa Shah. | Launched TweetAnalyzer for instant detection and visualization of Twitter trends. | Delivers practical insights from Twitter information for different uses such as job hunting and news alerts. | Restricted scalability for extremely large datasets; possible delays in real-time data processing. | Twitter API, visualization tools |
| Umit Demirbaga | created HTwitt, a Hadoop-centric platform designed for the analysis and visualization of streaming Twitter data. | Effectively addresses large data issues in tweet categorization; incorporates machine learning methods for enhanced precision. | Difficulty in setup and upkeep; necessitates expertise in the domain for best results | Hadoop ecosystem, MapReduce algorithms, Naive Bayes classifier |

1. **COMPARATIVE ANALYSIS**

**3.1 Performance of Techniques:**

The effectiveness of various models in toxic comment classification can be compared based on their performance metrics, particularly accuracy and F1-score. Below is a summary of the performance of different techniques:

| **Technique** | **Accuracy (%)** | **F1-Score (%)** | **Comments** |
| --- | --- | --- | --- |
| Logistic Regression | 78.5 | 75.4 | Provides a baseline for comparison; performs adequately on simpler datasets. |
| Support Vector Machine (SVM) | 82.3 | 80.1 | Effective in high-dimensional spaces; struggles with complex language patterns. |
| Decision Trees | 75.0 | 72.0 | Easy to interpret but often overfits to training data. |
| Random Forest | 85.0 | 83.5 | Robust against overfitting; performs well with imbalanced datasets. |
| Long Short-Term Memory (LSTM) | 89.2 | 88.0 | Excels in capturing sequential dependencies; highly effective for text data. |
| BERT (Bidirectional Encoder Representations from Transformers) | 92.5 | 91.3 | State-of-the-art performance; leverages contextual embeddings for superior accuracy. |

**3.2 Limitations and Drawbacks:** Despite the advancements in toxic comment classification techniques, several limitations have been identified across studies:

* **Generalization Issues**: Numerous models, particularly ones utilizing traditional machine learning methods, often experience overfitting to the training data, leading to subpar outcomes on unfamiliar data.
* **Handling Sarcasm and Nuance**: The ability to detect sarcasm, irony, and other nuanced expressions remains a challenge for most models, leading to misclassification of comments that require deeper contextual understanding.
* **Interpretability**: As models become increasingly complex, understanding their decision-making processes becomes difficult, which can hinder user trust and acceptance of automated moderation systems.

**3.3 Effectiveness of Different Techniques**

The effectiveness of various NLP and machine learning techniques in addressing the identified challenges is noteworthy:

* **Natural Language Processing Techniques**: Sentiment analysis and word embeddings greatly improve model performance by offering more comprehensive contextual details about comments.
* **Deep Learning Methods**: Neural networks, especially transformer-based models such as BERT, have shown outstanding ability in dealing with linguistic complexity and generalizing across various contexts by utilizing extensive data for learning.
* **Data Handling Strategies**: Techniques for managing data imbalance—such as oversampling minority classes or employing synthetic data generation—have shown promise in improving model robustness and ensuring balanced training datasets.

1. **DISCUSSIONS**

**4.1 Synthesis of Findings**

The comparative study of methods for classifying toxic comments uncovers several important insights that indicate present trends in the domain. Significantly, deep learning models, especially transformer-based architectures such as BERT, have surfaced as the most potent instruments for correctly identifying toxic comments. Their capacity to utilize contextual embeddings enables them to grasp the intricacies of language, such as sarcasm and subtle expressions that conventional models frequently overlook. This tendency highlights a move towards more advanced approaches that emphasize grasping the nuances of human interaction.

Additionally, the increasing acknowledgment of data imbalance as a significant issue has resulted in the creation of numerous data management techniques, including oversampling and synthetic data creation. These methods seek to develop more equitable datasets, thus enhancing model effectiveness and generalization abilities. Furthermore, the significance of model interpretability is becoming more recognized, as researchers are increasingly concentrating on approaches that improve clarity in automated decision-making systems. This transition is crucial for building user confidence in moderation systems, particularly as these technologies increasingly blend into online platforms.

**4.2 Key Gaps and Future Directions**

Although progress has been made in classifying toxic comments, there are still numerous gaps in the existing research field:

1. **Requirement for Improved Datasets**: Although current datasets such as Jigsaw and Wikipedia comments offer useful assets for model training, there is a necessity for wider and more representative datasets. Existing datasets might not sufficiently reflect the linguistic variety found among various cultures and online communities.
2. **Strong Management of Context**: Despite the advancements in context awareness by transformer models, there remains a necessity for models that can dynamically adjust to different contexts and cultural standards. Future studies ought to concentrate on creating context-sensitive models that can more effectively grasp the situational subtleties of toxicity.
3. **Enhanced Clarity**: The intricate nature of deep learning models frequently leads to a deficiency in interpretability, making it challenging for users to comprehend the decision-making process. There is an urgent requirement for studies aimed at improving model explainability to guarantee accountability in automated moderation systems.

**4.3 Potential Improvements:**

To address these gaps and enhance the effectiveness of toxic comment classification systems, several avenues for future research can be explored:

1. **Developing Context-Aware Models**: Future studies could investigate the creation of models that incorporate contextual information from user profiles or community guidelines to improve classification accuracy. Such models would be better equipped to discern acceptable language within specific contexts.
2. **Enhancing Interpretability**: Research should focus on integrating explainable AI techniques into toxic comment classification models. Methods such as LIME and SHAP can be further refined to provide clearer insights into model predictions, helping users understand why certain comments are classified as toxic or non-toxic.
3. **Utilizing Multilingual Datasets**: As online platforms operate globally, developing multilingual datasets will be crucial for creating robust classification systems that can handle diverse languages and cultural expressions. This approach will help ensure that toxic comment detection is effective across different linguistic contexts.
4. **CONCLUSION**

This review has examined the vital concern of classifying toxic comments on online platforms, highlighting the pressing necessity for efficient moderation tactics to cultivate more positive digital spaces. The rise in user-generated content has brought about significant challenges like the prevalence of hate speech, harassment, and other harmful behaviour that can negatively impact constructive conversations and overall community well-being. Researchers have made significant progress in developing models that can automatically identify and classify harmful comments using machine learning and natural language processing techniques.   
  
The comparative evaluation emphasized the enhanced capability of advanced deep learning models, especially transformer-based structures like BERT, which are proficient in grasping linguistic subtleties and contextual details. These models exhibit greater accuracy and F1-scores than conventional machine learning methods, highlighting the necessity of embracing advanced techniques to tackle the intricacies of language.  
  
In spite of these advancements, numerous gaps still exist in the present research environment. The necessity for more varied and representative datasets is crucial to guarantee that models can generalize in various contexts and cultures. Moreover, improving model interpretability is essential for building user confidence in automated moderation systems. Upcoming studies should concentrate on creating context-sensitive models that can adjust to changing norms and expectations within various online communities.  
  
This evaluation acts as a basis for upcoming studies and progress in the area of toxic comment classification. By tackling the recognized gaps and emphasizing innovative approaches, researchers can aid in developing more efficient systems that not only detect toxic behaviour but also foster respectful engagement among users. In the end, efficient moderation via automated systems has the great potential to protect user welfare, maintain the values of free speech, and foster inclusive online environments where various perspectives can flourish without the threat of harassment or abuse.

1. **REFERENCES**
2.  H. Rahul, J. Kajla, J. Hooda, and G. Saini, "Classification of Online Toxic Comments Using Machine Learning Algorithms," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2020, pp. 1119-1123. doi: 10.1109/ICICCS48265.2020.9120939.
3.  N. Garlapati, G. Malisetty, and G. Narayanan, "Classification of Toxicity in Comments using NLP and LSTM," in *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, 2022, pp. 16-21. doi: 10.1109/ICACCS54159.2022.9785067.
4.  A. Ashish, A. Rani, and H. Shyan, "A Comparative Study and Analysis on Toxic Comment Classification," in *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 2023, pp. 783-787. doi: 10.1109/ICSCSS57650.2023.10169771.
5.  F. Museng, A. Jessica, N. Wijaya, A. Anderies, and I. A. Iswanto, "Systematic Literature Review: Toxic Comment Classification," in *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, Yogyakarta, Indonesia, 2022, pp. 1-7. doi: 10.1109/ICITDA55840.2022.9971338.
6.  V. Swetha, R. Anuhya, E. S. Sowmya, and A. Geethanjali, "Building a Toxic Comments Classification Model," in *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2021, pp. 1519-1523. doi: 10.1109/ICECA52323.2021.9675911.
7.  M. Husnain, A. Khalid, and N. Shafi, "A Novel Preprocessing Technique for Toxic Comment Classification," in *2021 International Conference on Artificial Intelligence (ICAI)*, Islamabad, Pakistan, 2021, pp. 22-27. doi: 10.1109/ICAI52203.2021.9445252.
8.  K. Dubey, R. Nair, M. U. Khan, and P. S. Shaikh, "Toxic Comment Detection using LSTM," in *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, Bengaluru, India, 2020, pp. 1-8. doi: 10.1109/ICAECC50550.2020.9339521.
9.  J. Zhong, X. Zhang, and C. Jiang, "Improving Toxic Comment Classification via Data Augmentation and Transfer Learning," in *2021 IEEE International Conference on Data Mining (ICDM)*, Auckland, New Zealand, 2021, pp. 1045-1050. doi: 10.1109/ICDM.2021.00133.
10.  S. M. Agrawal and A. Narang, "Detection of Hate Speech and Offensive Language on Twitter using Ensemble Deep Learning Models," in *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, 2021, pp. 93-98. doi: 10.1109/AIMV52357.2021.9463080.
11.  E. Wulczyn, N. Thain, and L. Dixon, "Ex Machina: Personal Attacks Seen at Scale," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1391-1399. doi: 10.1145/3038912.3052591.
12.  S. Jain, M. Verma, and R. Rastogi, "Sentiment Analysis and Toxicity Classification on Twitter Data," in *2022 IEEE 7th International Conference on Computing, Communication and Security (ICCCS)*, 2022, pp. 1-6. doi: 10.1109/ICCCS54332.2022.9789835.
13.  L. Song, R. Y. Lau, and C. Yin, "Discriminative Topic Mining for Social Spam Detection," in *PACIS*, 2014, p. 378.
14.  A. Singh, P. Raj, and S. P. Mishra, "Deep Learning-Based System for Toxic Comment Classification in Social Media," in *2021 IEEE 4th International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India, 2021, pp. 1-6. doi: 10.1109/ISCON52037.2021.9702367.
15.  M. Naseem, A. Razzak, M. M. Khan, and M. Prasad, "Toxic Comment Classification Using Deep Learning Techniques," in *2021 International Conference on Computing and Communication Technologies (ICCCT)*, 2021, pp. 17-21. doi: 10.1109/ICCCT53315.2021.9677614.
16.  G. S. Brar and A. S. Sharma, "Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques," unpublished.