Toxic Comments Classification

Sakshi Kaurase[1], Sayukta Giradkar[2], Sanchi Borkar[3], Bhumika Rasekar[4], Prof.Rupatai Lichode[5]
Students, Department of Computer Science and Engineering[1,2,3,4]
Assistant Professor, Department of Computer Science Engineering[5]
Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, Maharashtra, India

Abstract*—Nowadays, insulting remarks are frequently uttered by the same individuals. Not just in offline contexts. However, there are also online ecosystems like social networking sites and online platforms. Therefore, identifications is integrated with security mechanisms on all websites and social networks applications including all communities in the digital world. Any undesirable internet behavior needs to be handled. This should urge the defense sector to respond accordingly. The objective of this study is to analyze ,understanding messages to toxicity, such as Threat, Insult and Identity based hatred. This is accomplished using a Jigsaw- created dataset of labeled Wikipedia comments. Machine learning model is created and trained.

Keyword*—Toxicity, Threat, Insult, Hatred, Machine learning

I. INTRODUCTION

These days , People can express themselves and their thoughts, as well as engage in discussions with others, using these platforms. In such a situation, it is understandable that disagreements may occur. However, these conversations frequently take a dark turn, resulting in social media conflicts in which one side may use harsh language known as toxic comments. These poisonous statements might be threatening, disrespectful, or motivated by hatred of a certain identity. As a result, these increase the risk of online abuse and harassment. Detecting toxic comments has been a significant challenge for all scholars working in the field of research and development. To address the above mentioned issue, we created a Toxic Comment Classification System utilizing Machine Learning. The system is intended to detect and classify harmful comments or words when chatting. It promotes good communication among users and discourages the use of unpleasant , obscene ,threats or hatred words. Manually filtering the large amount of online content is impractical, making classification an essential duty. Automatically categorise comments as harmful or non-toxic using machine learning models, including deep learning and NLP approaches. The complexity of human language, including context, sarcasm, and cultural differences, might impact how toxicity is communicated and perceived. Developing effective toxic comment classification systems is vital for creating safe and inclusive online environments.

II. LITERATURE SURVEY

This is a review of recent research in machine learning-based harmful comment classification, emphasizing datasets, approaches, and difficulties:

1. Conventional Methods of Machine Learning (2012-2016)

Data sets: Small, specially gathered datasets from forums and social media sites like Twitter were employed in early research. Methods: Conventional models such as Logistic Regression, Support Vector Machines (SVM), and Naive Bayes were frequently employed. For feature extraction, these models mostly used bag-of-words and TF-IDF.

Challenges: The models have trouble recognizing subtleties and context, including indirect communication, irony, and cultural differences. The model's effectiveness and generalizability were restricted by limited dataset sizes and the absence of labeled data. Davidson et al.

Key Study: To illustrate the difficulties in differentiating hate speech from ordinary offensive content, Davidson et al. (2017) produced one of the earliest publicly accessible hate speech datasets and classified offensive language using SVM and Logistic Regression.

2. Large-scale datasets and deep learning models (2017-2018)

Datasets: The Jigsaw Toxic Comment Classification Challenge dataset, which was made available on Kaggle in 2017 and included labels for a number of toxic categories such as hate speech, insults, and threats, was a significant breakthrough. *Methods:* Because of their ability to recognize intricate patterns in text, deep learning models—in particular, convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—have gained popularity. For improved semantic comprehension, models were trained using embeddings such as Word2Vec and GloVe. *Challenges:* Deep learning models were difficult to interpret and demanded a lot of processing power, even with advancements. Because there were more non-toxic than toxic samples in the dataset, class imbalance also resulted in skewed predictions. Important Research: Zhang et al. (2018) used LSTMs and CNNs on the Jigsaw dataset, achieving better accuracy than traditional models but identifying interpretability as a key challenge in deploying these systems in real-world applications.

3. Contextual Embeddings and Transformers (2019–2021)

Datasets: More thorough datasets with an emphasis on identifying bias and guaranteeing equitable classification across demographics surfaced, such as the Jigsaw Unintended Bias in Toxicity Classification dataset.

Methods: Because transformer-based models such as BERT, RoBERT, and Distil BERT can capture contextual relationships in language, they have gained widespread adoption and greatly improved performance when managing subtle poisonous comments.

Challenges: Because models frequently mirrored biases in the data, problems with model bias and fairness emerged. To combat these biases, methods such as adversarial training were developed; nevertheless, this increased complexity and did not completely address issues of fairness.

Key Study: Park et al. (2020) demonstrated that contextual embeddings enhanced accuracy but did not completely eradicate model bias, especially for underrepresented demographic groups, by utilizing BERT in conjunction with adversarial training to reduce bias.

4. Multilingual Toxicity Detection and Fairness Concerns(2022-Present)

Datasets: To address the need to detect toxicity in a variety of linguistic and cultural contexts, multilingual toxicity detection datasets, including the Jigsaw Multilingual Toxicity Classification dataset, offered labeled data across languages.

Methods: For managing linguistic variation, multilingual transformers such as XLM-R and mBERT are well-liked options. To further enhance performance across languages, ensemble strategies that combine different models have also been attempted.

Challenges: Different cultural contexts and translation accuracy are problems for cross-lingual models. Since definitions of toxicity can vary throughout languages and cultures, fairness and ethical issues are important since they cast doubt on the universality of classification schemes.

key Study: Chaudhuri et al. (2023) employed mBERT and XLM-R to study multilingual toxicity detection. They showed that while multilingual models increase inclusivity, they also need cultural adaptation to detect toxicity in a meaningful and equitable way.

III. METHODOLOGIES

1. *Gathering and preparing datasets* : The first step is to gather a dataset of both labeled and non-sub-toxic comments. This entails gathering information from pertinent platforms and sources. Preprocess the acquired text data by removing stopwords, punctuation, and special characters. To further standardize the format, tokenize the text and change its capitalization to lowercase.

2. *Deep learning feature extraction*: Use deep learning methods to glean valuable features from the text material that has already been preprocessed. To identify complex patterns and representations in the text, build a deep learning architecture, such as a neural network with convolutional and recurrent layers. The following categorization challenge will use these attributes as inputs.

3. *Support Vector Machine (SVM) classifier integration:* Include an SVM classifier in the deep learning framework. SVMs can supplement the feature representation that the deep model has

acquired and are well known for their effectiveness in binary classification problems. Through this integration, the advantages of SVMs and deep learning are combined to increase classification accuracy.

4. *Model training and optimization:* Use the labeled dataset to train the combined deep learning with SVM model. Optimize the model's performance by adjusting its hyperparameters using methods like grid search and cross-validation. The model is guaranteed to produce the best categorization results thanks to this iterative method.

5. *Evaluation and validation*: Assess the trained model's performance using appropriate measures like as accuracy, precision, recall, and F1-score. Validate the model's performance by running it over a different test dataset and evaluating its ability to appropriately classify subtoxic remarks. This stage verifies that the model generalizes well to unseen data and accurately identifies sub-toxic language.

6. *Maintenance* : Maintenance for sub toxic comment classification using deep learning with SVM involves continuous data updates to keep the model relevant, regular performance monitoring and retraining to adapt to evolving language patterns, periodic hyperparameter tuning for optimization, proactive measures to address concept drift, and implementation of robust security and privacy measures to safeguard the model and data.

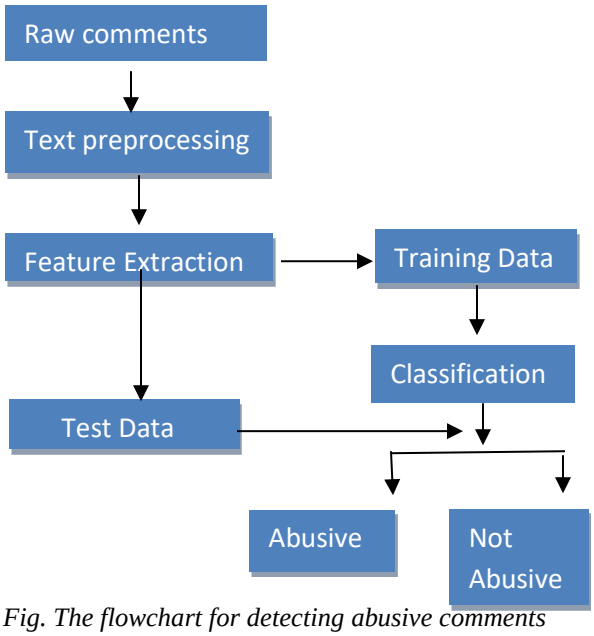


Fig. The flowchart for detecting abusive comments

IV. CHALLENGES IN TOXIC COMMENTS CLASSIFICATION

1. Deepfake and Misinformation: Because context and trustworthiness might be weakened by the increasing number of deepfake technology and manipulated content, it may be more difficult to identify harmful remarks.

2. Increasing Polarization: Discussions on the internet have gotten progressively divisive, with extreme viewpoints reflected in comments that can make it hard to distinguish between toxic and non-toxic substances, making classification more difficult.

3. Emerging Platforms: Models have to adapt to a range of user preferences as a result of the introduction of innovative formats and methods of interaction by new social media platforms and communication tools.

4. Developments in NLP: While advances in natural language processing improve classification, they also increase the bar for accuracy, imposing pressure on models to remain up to date.

5. Multimodal Content: Classification systems have new challenges in recognizing toxicity in a multimodal setting when platforms start to include images, videos, and audio in addition to text.

6. Adaptation to User Behavior: Users frequently modify their language to prevent detection (e.g., by adopting memes or coded language), necessitating ongoing model changes.

7. Toxicity Amplification: Because toxic comments can spread quickly due to their widely nature, it can be challenging to efficiently monitor and categorize them in real time.

8. Implications for Ethics: The necessity for models that are not just accurate but also just and equitable is highlighted by the ongoing debates around bias and fairness in AI.

9. Data privacy issues: Attention to data protection laws (such as GDPR) restricts the amount of data that

can be used for training, which affects the resilience and performance of the model.

10. Community Standards Variation: It can be tough to develop a classification system that works for all online communities because they have different norms and standards for appropriate conduct.

V. APPLICATIONS

Several websites utilize toxic comment classification to maintain polite and safe online environments. Here are a few basic samples of its application:

1. *Social Media Sites:*It assists social media sites like Instagram, Twitter, and Facebook in automatically identifying and removing offensive or dangerous remarks. People can now share on social media more freely and without fear of harassment or violence.

2. Blogs and new websites: Blogs and Websites for News Comment sections are available on a lot of news websites. These websites can filter out nasty or unsuitable remarks with the use of toxic comment classification. This prevents conflicts and hate speech and maintains the discussions civil.

*3. Playing video games online:*Toxic conduct in online games, such as insults or name-calling, can make the experience for other players miserable. Players that use foul language can be instantly muted or banned via toxic comment detection.

4. Client Assistance: Those who communicate with customer service via emails or chat rooms occasionally use inappropriate language. This system has the ability to block very offensive messages entirely or mark them so that human agents can handle them with additional caution.

5. Online Learning Environments: People can leave comments or pose questions on online learning systems. It is ensured that no one is harassed or treated disrespectfully in these areas by toxic remark detection. This promotes a secure and encouraging learning atmosphere.

*6. Discussion Boards:*On websites such as Reddit or Quora, users post questions and receive answers. The use of toxic remark detection ensures that conversations remain civil and avoid becoming combative. More people are encouraged to join in conversations as a result of not having to worry about being attacked.

7.A stopping cyberbullying: By identifying harmful remarks early on, toxic comment detection can assist in stopping cyberbullying, particularly on websites frequented by children and teenagers. This encourages safer interactions and lessens the risk of emotional injury.

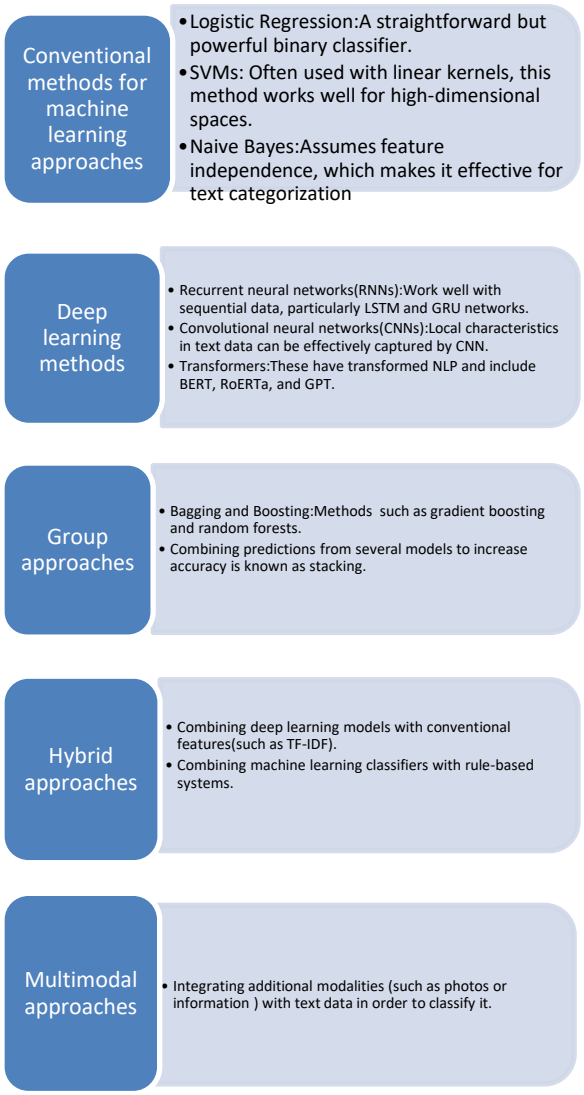
8.Workplace chats: Office Conversations Employees at businesses communicate using email or apps like Slack. The identification of toxic comments can identify language that is improper or unprofessional. This maintains a polite and professional work environment.

9.Monitoring Hate Speech: With the use of this technology, law enforcement may monitor and suppress hate speech and threats made online, averting harm to people in the real world. By doing this, communities are protected against harmful or bigoted behaviors.

Toxic comment classification in each one of these locations contributes to a broader and secure online environment for all users.

VI. COMPARITIVE ANALYSIS OF DIFFERENT METHODS

The particular needs of the application, such as the amount of the dataset, the available computing power, and the intended trade-off between interpretability and performance, will determine which poisonous comment classification approach is best. While deep learning techniques are favored for intricate and extensive applications, traditional methods could still be appropriate for fewer datasets or simpler jobs. Particularly in real-world situations where data volatility and noise are prevalent, ensemble and hybrid approaches might offer a suitable balance.



VII. CONCLUSION

In conclusion, machine learning-based hazardous comment classification has demonstrated great promise for improving the safety and security of online conversation. Researchers have created reliable systems to identify and reduce harmful information by utilizing a variety of machine learning techniques, including both conventional models like logistic regression and more sophisticated models like deep neural networks and transformer-based architectures. These models have proven to be highly accurate and successful in managing a variety of complicated language inputs, especially when paired with feature engineering and data pretreatment techniques. There are still difficulties in spite of these developments. The effectiveness and generalizability of the model are hampered by problems like context sensitivity, data imbalance, and the nuances of harsh language. Furthermore, automated moderation's impartiality and inclusivity are at danger due to ethical issues, such as possible biases in training data. In order to guarantee objective categorization, future research should concentrate on creating increasingly complex models that comprehend complex language and enhancing interpretability and fairness. To make these models worldwide applicable, it is also necessary to investigate the classification of hazardous comments across languages and cultures. All things considered, even if machine learning-based toxic comment classification has advanced significantly, more study and improvement are necessary to create online spaces that are safer and more civil.

REFERENCES

[1] K. Dubey, R. Nair, M. U. Khan and P. S. Shaikh, "Toxic Comment Detection using LSTM," 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC), Bengaluru, India, 2020, pp. 1-8, doi: 10.1109/ICAECC50550.2020.9339521.

[2]Rahul,H. Kajla, J. Hooda and G. Saini, "Classification of Online Toxic Comments Using Machine Learning Algorithms," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1119-1123, doi: 10.1109/ICICCS48265.2020.9120939.

[3] V. Swetha, R. Anuhya, E. S. Sowmya and A. Geethanjali, "Building a Toxic Comments Classification Model," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2021, pp. 1519-1523, doi: 10.1109/ICECA52323.2021.9675911.

[4] F. Museng, A. Jessica, N. Wijaya, A. Anderies and I. A. Iswanto, "Systematic Literature Review: Toxic Comment Classification," 2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA), Yogyakarta, Indonesia, 2022, pp. 1-7, doi: 10.1109/ICITDA55840.2022.9971338.

[5] Gladwin, E. V. Renjiro, B. Valerian, I. S. Edbert

and D. Suhartono, "Toxic Comment Identification and Classification using BERT and SVM," 2022 8th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICST56971.2022.10136295.

[6] Rahul et al. “Classification of Online Toxic Comments Using Machine Learning Algorithms.” 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (2020): 1119-1123.

[7] Sonawane, Chaitanya & Bagale, Tejaswini & Kawade, Preeti & Ogale, Swarada & Singru, Prof. (2024). TOXIC COMMENT CLASSIFICATION SYSTEM USING DEEP LEARNING. IJARCCE. 13. 10.17148/IJARCCE.2024.13465.