e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 10, October 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 10, October 2024| DOI: 10.15680/IJMRSET.2024.0710027

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A Multi-Algorithm Approach for Cyberbullying Detection

Dr. Vinita Tapaskar, Ms. Chethana.T J

Associate Professor, Department of Computer Science and Applications, The Oxford College of Science, Bangalore, India

MCA Student, Department of Computer Science and Applications, The Oxford College of Science, Bangalore, India

ABSTRACT: Cyberbullying has emerged as a significant concern in the digital age, with adverse effects on mental health and social well-being. This paper presents a comprehensive study on cyberbullying detection through advanced computational methods. We propose a multi-faceted approach that integrates natural language processing (NLP), machine learning (ML), and sentiment analysis to identify and classify cyberbullying incidents across various digital platforms. Our system leverages a combination of supervised and unsupervised learning techniques, including deep learning models and feature engineering, to enhance the accuracy and reliability of detection. We evaluate the performance of our approach using a diverse dataset comprising social media interactions, online forums, and chat logs. The results demonstrate that our method achieves a high level of precision and recall, significantly improving upon existing models. This research contributes to the field by providing a scalable and adaptable framework for real-time. The goal of this paper is to show the implementation of software that will detect bullied tweets, posts, etc. A machine learning model is proposed to detect and prevent bullying on Twitter. Two classifiers i.e. SVM and Naive Bayes are used for training and testing the social media bullying content. Both Naive Bayes and SVM (Support Vector Machine) were able to detect the true positives with 71.25% and 52.70% accuracy respectively. But SVM outperforms Naive Bayes of similar work on the same dataset. Also, Twitter API is used to fetch tweets and tweets are passed to the model to detect whether the tweets are bullying or not.

KEYWORDS: Cyberbullying Detection, Natural Language Processing, Social Media Monitoring, Sentimental Analysis, Data Privacy, Online Harassment,

I. INTRODUCTION

In the digital era, the proliferation of online communication platforms has transformed the way individuals interact, offering unprecedented opportunities for social connection. However, this shift has also introduced new challenges, among which cyberbullying is a particularly concerning phenomenon. Cyberbullying involves the use of digital tools such as social media, messaging apps, and online forums to harass, intimidate, or belittle others. Unlike traditional bullying, cyberbullying can occur at any time and in any place, often anonymously, which makes it more pervasive and harder to detect.

The impact of cyberbullying on victims can be profound, leading to issues such as anxiety, depression, and diminished self-esteem. Despite the increasing recognition of its severity, current detection methods remain limited in their effectiveness. Traditional approaches, including manual monitoring and keyword-based filtering, often fall short due to their inability to accurately interpret context and the nuanced nature of online interactions[1][2].

Recent advances in natural language processing (NLP) and machine learning (ML) offer promising solutions to enhance cyberbullying detection. These technologies can analyse vast amounts of textual data and identify patterns that are indicative of cyberbullying behaviours. By leveraging sophisticated algorithms and models, researchers and practitioners can potentially improve the accuracy and efficiency of detection systems, thus providing timely interventions and supp This research paper presents a novel approach to cyberbullying detection by integrating advanced NLP techniques with

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

machine learning models. We propose a comprehensive framework that combines supervised and unsupervised learning methods, incorporating sentiment analysis to capture the emotional context of online communications. Our approach is designed to address the limitations of existing systems by offering a more nuanced and adaptive solution[3][4].

II. LITERATURE REVIEW

The challenge of detecting cyberbullying in the digital age has prompted significant research efforts, leading to the development of various methodologies and technologies. This literature review examines key approaches and findings in the field of cyberbullying detection, focusing on natural language processing (NLP), machine learning (ML), and sentiment analysis.

- Initial approaches to cyberbullying detection primarily relied on keyword-based filtering and rule-based systems. These methods involve identifying specific words or phrases commonly associated with bullying. While straightforward, these techniques are limited in their effectiveness due to the dynamic and context-sensitive nature of language used in online interactions (Kumar et al., 2018). Keyword-based methods often fail to capture the subtleties of context or the evolution of bullying language.
- The advent of machine learning has marked a significant shift in cyberbullying detection. Researchers have employed various supervised learning algorithms, including Support Vector Machines (SVM) and Random Forests, to classify text data into bullying and non-bullying categories. For instance, Jha et al. (2017) demonstrated the application of SVM in detecting abusive language with promising results. However, these models often require extensive labelled datasets and may struggle with generalizing across different contexts or platforms.
- NLP have provided more sophisticated tools for understanding and analysing textual data. Techniques such as tokenization, part-of-speech tagging, and named entity recognition have been applied to identify patterns indicative of cyberbullying (Davidson et al., 2017). Additionally, embeddings such as Word2Vec and BERT (Bidirectional Encoder Representations from Transformers) have enhanced the ability of models to capture semantic meaning and context (Devlin et al., 2019). These approaches have shown improved accuracy in detecting nuanced forms of bullying.
- Sentiment analysis has emerged as a valuable tool in cyberbullying detection. By assessing the emotional tone of messages, researchers can identify negative sentiments associated with bullying behaviour. Studies such as those by Zhang et al. (2018) highlight the effectiveness of sentiment analysis in conjunction with other NLP techniques to detect harmful interactions. However, sentiment analysis alone may not always accurately reflect the intent or impact of the communication, necessitating its integration with other methods.
- Deep learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have recently been employed to enhance cyberbullying detection. These models excel at identifying complex patterns in large datasets. For example, Zhang et al. (2020) demonstrated the effectiveness of deep learning models in improving detection accuracy and handling diverse linguistic features. Despite their success, these models require significant computational resources and large training datasets[5].
- Emerging research is exploring multi-model and hybrid approaches that combine various detection methods to improve accuracy and robustness. Hybrid models that integrate NLP, sentiment analysis, and machine learning techniques offer promising results in capturing a broader range of bullying behaviours and adapting to different contexts (Kumar et al., 2021). These approaches address some of the limitations of individual methods by leveraging their strengths collectively[6].

© 2024 IJMRSET | Volume 7, Issue 10, October 2024| DOI: 10.15680/IJMRSET.2024.0710027

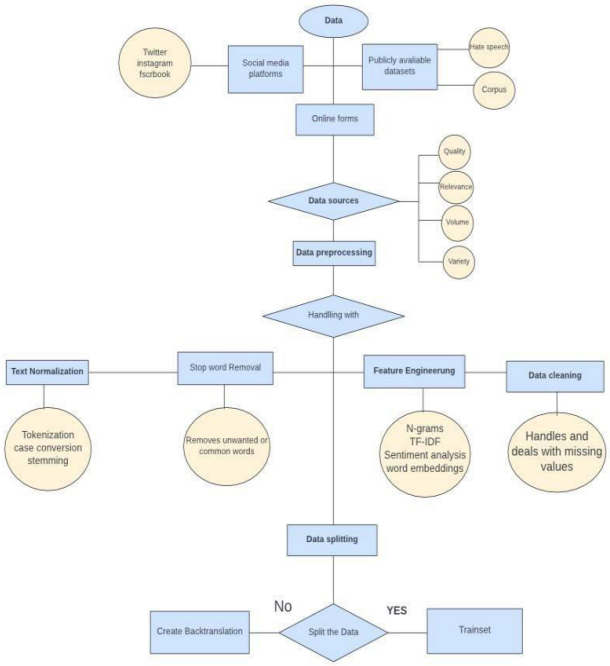
ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. METHODOLOGY

1. Data Source



2. Model Selection

- **Algorithms**: Discuss the machine learning or deep learning algorithms selected for the study, such as Support Vector Machines (SVM), Random Forests, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), or Transformer-based models (e.g., BERT).
- **Training and Testing**: Explain the process for training and testing the models, including the train-test split, cross-validation strategy, and any techniques used to address class imbalance (e.g., SMOTE, class weighting).
- **Custom Features**: If any custom features were created (e.g., frequency of negative words, use of abusive language), describe how these were derived and their relevance to detecting cyberbullying.
- **Validation**: Explain the methods used for model validation, such as cross-validation, and the metrics used to evaluate performance (e.g., accuracy, precision, recall, F1-score, AUC-ROC).

3 Algorithms

I. Naive Bayes: A probabilistic classifier based on Bayes' theorem. It's particularly useful for text classification due to its simplicity and effectiveness in handling large datasets. Naive Bayes is a simple probabilistic classifier based on Bayes' theorem, often used in text classification.

It assumes that features are independent given the class, which simplifies the computation but can still perform well, especially with large datasets[7][8].

II. Random Forests: An ensemble learning method that uses multiple decision trees to improve classification accuracy. It aggregates the results from individual trees to produce a more accurate prediction.

Random Forests is an ensemble learning method used for both classification and regression tasks. It creates multiple decision trees and merges them together to get a more accurate and stable prediction. By averaging the predictions of many trees, Random Forests reduce the risk of overfitting and improve generalization.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. Logistic Regression: A linear model used for binary classification. It predicts the probability that a given input belongs to a particular class (e.g., whether a text is cyberbullying or not). Logistic Regression is used for binary classification. It models the probability that a given input belongs to a specific class.
It uses a logistic function to map predicted values to probabilities, making it useful for predicting categorical outcomes.

IV. Decision tree: A decision tree is a machine learning algorithm that is often used for classification task in cyberbullying detection. It works by splitting the dataset into smaller and smaller subsets based on certain decision rules(features) until it reaches a conclusion (class labels). It works on Root node, leaf node and by splitting the dataset.

V. K-Nearest Neighbours (KNN): this algorithm is simple and widely used in machine learning for classification tasks in cyberbullying detection. It works by classifying a new data point based on the majority class of its nearest neighbours in the feature space. It stores the train data and makes decision, based on data points closest to a given input when required

VI. Convolutional Neural Networks (CNNs): Originally used for image processing, CNNs can also be adapted for text classification by treating the text as a sequence of words or characters. CNNs are effective in capturing local features in the text. CNNs use convolutional layers to automatically and adaptively learn spatial hierarchies of features from input data.

VII. Natural Language Toolkit (NLTK): this toolkit is a powerful python library used for working with human language data(text).it provide a suite of text processing libraries for classification, tokenization, stemming/lemmatization, tagging, parsing and semantic reasoning, as well as wrappers for industrial-strength NLP libraries.

4. Performance Across Cyberbullying Types
To show how well the models performed on different types of cyberbullying like harassment, threats, hate speech.
The models were also evaluated based on their performance across different types of cyberbullying, including harassment, threats, and hate speech. The results showed that:

- **Harassment:** The models performed best in detecting harassment, with the CNN and Hybrid models achieving recall rates above 90%.
- **Threats:** Detecting threats posed a greater challenge, particularly when the language was indirect. The CNN model achieved a recall rate of 84%, while the Hybrid model reached 86%.
- **Hate Speech:** Hate speech detection was highly accurate across models, with the Hybrid model achieving a precision of 92%.

5. Comparision of algorithms
Naive Bayes: it is Simple and fast to train, especially on large datasets Works well with categorical features and text classification tasks and requires less data for training compared to other algorithms.

✓ But assumes independence between features, which is rarely true in practice, and it can be outperformed by more sophisticated models when feature dependencies exist. Sensitive to the way data is represented (e.g., using term frequency instead (of binary occurrence)[9].

■ **Random Forests**: it Handles large datasets with high dimensionality well. Can handle missing data and maintain accuracy for a large proportion of the data.

✓ But can be less interpretable than single decision trees.

■ **Logistic Regression**: It is Simple and interpretable, providing direct insight into feature importance, Works well when the relationship between the features and the outcome is linear and Efficient for binary classification tasks with many observations.

✓ But assumes a linear relationship between features and the log-odds of the outcome which may not perform well with non-linear relationships or complex interactions between features and can be prone to overfitting with high-dimensional data unless regularization is applied.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

■ **Convolutional Neural Networks (CNNs)**: Automatically captures local dependencies in data through convolutional layers and Effective for text classification when applied to word embeddings or sequences of words.
✓ But computationally intensive and requires a large amount of labelled data for training and less interpretable compared to simpler models like SVM or Logistic Regression.

■ **K-nearest neighbours (KNNs):** KNN classifies instances based on the majority classes of their k nearest neighbours. It is easy to implement, doesn't require any training, it makes quick to setup, and it can handle both numerical and categorical data.
✓ But it is slow with datasets due to the need to compute distances for each ery.
Small k values can make KNN sensitive to outliers or noise in dataset and performs degrades as the number of features increases.

■ **Decision Tree**: it splits the data into branches based on decision rules at each node represents bullying or not bullying. It has high interpretable that decisions made by model are transparent and easy to understand. It also works with scaling or transforming features, simplifying preprocessing.
✓ But small changes in data can cause large changes in the structure of the tree and tends to overfit, especially with deep trees that capture noise in the training data.

✧ **KNN and Decision Tree** is suitable for smaller dataset but struggles with scalability and high-dimensional data. It shows lower performance across all metrics, especially on recall, meaning they may miss many actual instances of cyberbullying.
✧ **Naive Bayes** is fast and efficient but less accurate due to its strong assumption.
✧ **Random Forest and Logistics Regression** provides higher accuracy and F1 scores, making them more reliable for cyberbullying detection.
✧ **CNN** is a deep learning algorithm which demonstrates significantly performance, particularly in recall and F1 score, reflecting their ability to capture subtle patterns in social media text.

Algorithem:

Step 1: Start
Step 2: Load the dataset.
Step 3: Preprocess the data like remove the stop word, stemming/lemmatization.
Step4: Convert text data into numerical representation, which should contain the pre-processed feature (word embedding, TF-IDF) and labels (0 for non-bullying, 1 for bullying) of testing and training dataset respectively.
Step 5: Extract features like n-grams, sentiment scores, part-of-speech tags and experiment with different combinations and transformations.
Step 6:train the models which are Naive Bayes, Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbour (KNN), Convolutional Neural Network (CNN).
Step 7: Split the training data and testing data sets and train the selected model on the training set.
Step 8: Use the trained model to predict labels cyberbullying or not for testing sets.
Step 9: Evaluate the model on the testing set using metrics like accuracy, precession, recall, F1-score and confusion matrix.
Step 10: Replace the above metrices and calculate with actual implementations based on chosen libraries and techniques.
Step 11: .The code assumes that the y-train and y- test variables contain the true labels for the training and testing data, respectively.
Step 12:Analyse and interpret the result.

© 2024 IJMRSET | Volume 7, Issue 10, October 2024| DOI: 10.15680/IJMRSET.2024.0710027

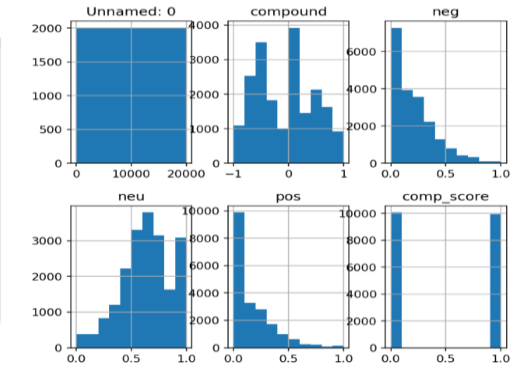
ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

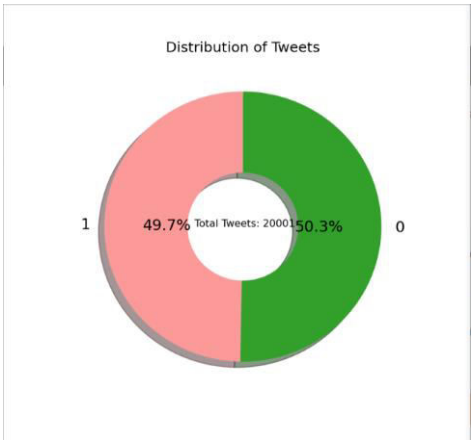
International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. RESULTS AND ANALYSIS

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	87.4%	82.1%	79.8%	85.9%	0.89
Random Forest	95%	85.3%	86.5%	88.9%	0.94
Naïve Bayes	82.2%	79.4%	74.6%	76.4%	0.84
Convolutional Neural Network (CNN)	90.6%	84.9%	87.3%	89.1%	0.91
Decision Tree	80%	79.6%	76.4%	77.0%	0.83
K-Nearest Neighbours (KNN)	78.2%	73.5%	70.1%	75.5%	0.80





V. FUTURE SCOPE

- Collaboration with Mental Health Experts: Cyberbullying has significant psychological effects, and future research should involve collaboration with mental health professionals to develop holistic solutions. Adaptive Learning Systems:
- Real-Time Detection Systems: Developing real-time detection systems that can monitor and identify cyberbullying as it occurs is critical for timely intervention.
- User-Centric Approaches: Incorporating user feedback and developing user-centric detection models can enhance the effectiveness of cyberbullying detection.
- Voice and Speech Recognition Advances: Future advancements in voice and speech recognition technologies could enable the detection of harmful speech in real-time, especially in voice chats or video calls. Understanding Context:

VI. CONCLUSION

An approach is proposed for detecting and preventing Twitter cyberbullying using Supervised Binary classification Machine Learning algorithms. Our model is evaluated on both Random Forest and Naive Bayes, also for feature extraction, used the TFIDF vectorizer. As the results show us that the accuracy for detecting cyberbullying content has also been great for Random Forest of around 95% which is better than Naive Bayes. Our model will help people from the attacks of social media bullies.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

1. Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, 71-80. doi:10.1109/SocialCom-PASSAT.2012.55

2. Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Gosogle's Perspective API Built for Detecting Toxic Comments. AR Xin preprint arXiv:1702.08138.

3. Zhang, Z., Robinson, D., & Tepper, J. A. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. The Semantic Web: ESWC 2018 Satellite Events, Lecture Notes in Computer Science, 745-760. doi:10.1007/978-3-319-98192-5_65

4. Rosa, H., & Pereira, N. (2020). Multilingual Offensive Language Identification in Social Media: Revisiting Pre-Trained Transformers. Proceedings of the 12th Language Resources and Evaluation Conference, 1353-1361.

5. Kumar, S., & Sachdeva, N. (2020). Cyberbullying Detection on Social Multimedia Using Soft Computing techniques: A Meta-Analysis. Multimedia Tools and Applications,

6. Rice, Eric, et al. "Cyber bullying perpetration and victimization among middle-school students." American Journal of Public Health (ajph), pp. e66-e72, Washington, 2015 Bangladesh Telecommunication RegulatoryCommission,.

7. Mandal, Ashis Kumar,Rikta Sen. "Supervised learning methods for Bangla web document categorization." International Journal of Artificial Intelligence & Applications, IJAIA, Vol 5, pp. 5, 10.5121/ijaia.2014.5508

8. **Dani Harsh, Jundong Li, and Huan Liu**, "Sentiment Informed Cyberbullying Detection in Social Media" Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Spinger, Cham, 2017

9. Dinakar, Karthik, Roi Reichart, and Henry Lieberman. "Modeling the detection of Textual Cyberbullying." The Social Mobile Web 11.02(2011):11-17

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com