iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 Issue: XII Month of publication: December 2023

DOI: https://doi.org/10.22214/ijraset.2023.57629

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XII Dec 2023- Available at www.ijraset.com

Automated Detection and Symbolic Replacement of Abusive Language using Deep Learning in Online Platforms

Simran Dhankar[1], Utkarsh Jain[2], Vansh Bansal[3], Prof. Naved Ahmad[4]
[1, 2, 3]ADGITM, CSE Department

Abstract: This research addresses the pressing challenge of curbing abusive language in online platforms through the implementation of advanced deep learning techniques. Focused on Natural Language Processing (NLP), this study aims to develop a robust automated censorship system capable of swiftly detecting and mitigating abusive content. By leveraging the prowess of deep learning algorithms, particularly in neural network architectures, the proposed system aims to proactively identify and censor abusive language across various online platforms. Key components involve training models to comprehend contextual nuances, enabling accurate recognition of abusive language patterns. Through this approach, the research aims to significantly contribute to online moderation mechanisms, ensuring a safer and more respectful online environment. The integration of deep learning methodologies within automated censorship systems represents a pivotal step towards mitigating the spread of abusive language, thereby fostering healthier and more inclusive online communities.

Keywords: Abusive Language, Deep Learning, Online Moderation, Natural Language Processing, Automated Censorship

I. INTRODUCTION

The dynamic landscape of online communication, dominated by the colossal presence of platforms like Twitter, Facebook, and Instagram, has transformed digital interaction. However, this paradigm shift has also unveiled a disconcerting trend—a pervasive surge in abusive language within these online spaces. The unrestricted nature of communication channels has precipitated an alarming occurrence of uncontrolled discourse, marked by the frequent use of derogatory and offensive language, often including abusive words and phrases.

The exponential growth of online social networks and microblogging sites, synonymous with the burgeoning era of big data, presents a compelling domain for research. This realm encapsulates an array of diverse backgrounds, cultures, and interests, facilitating a rich tapestry of user-generated content. Yet, within this tapestry lies a concerning pattern—the unbridled propagation of abusive language.

Previous studies have underscored the dearth of effective mechanisms to filter out abusive language, attributing its prevalence to the absence of robust tools for content moderation and a lack of empathy among users. Children and adolescents, particularly susceptible to the influence of online content, face the risk of imbibing abusive language from these platforms.

Manually filtering abusive language within the vast expanse of social media platforms becomes an insurmountable challenge owing to the sheer volume of users engaging in such discourse. This necessitates the implementation of automated systems capable of discerning and replacing abusive language in real-time.

The research outlined in this paper aims to delve into this critical issue by exploring the domain of Machine Learning. Specifically, the focus lies on employing the Random Forest Algorithm to detect abusive language within the corpus of customer reviews. The overarching goal is to identify an adept classifier to detect abusive content accurately, primarily aimed at online blogging sites and social networks.

The necessity to detect and censor abusive language on these platforms has become imperative, given the monumental scale at which they operate. Traditional mechanisms, such as human moderation and the implementation of regular expressions and blacklists, fall short in the face of the vastness and dynamism of these platforms.

This research aims to bridge the existing gaps in prior literature, which lacks cohesion due to disparate methodologies dispersed across various fields, including AI, NLP, and Web Sciences. By consolidating these diverse strands of research, the objective is to develop a sophisticated and state-of-the-art method for detecting abusive language in user comments.

II. LITERATURE REVIEW

The surge in user-generated content on social media platforms has sparked a concomitant rise in abusive language, necessitating effective automated detection and mitigation mechanisms. Existing literature extensively investigates methodologies integrating deep learning into Natural Language Processing (NLP) for combating abusive language online.

Early works by Davidson et al. (2017) and Nobata et al. (2016) underscore the challenges of detecting nuanced abusive language, highlighting the need for context-aware models. Recent studies, such as Fortuna and Nunes (2020) and Zhang et al. (2021), emphasize the role of deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), in discerning contextual cues and linguistic nuances in abusive language detection.

Furthermore, advanced techniques incorporating Transformer-based architectures, as explored by Vaswani et al. (2017) and Devlin et al. (2019), exhibit promising results in capturing intricate linguistic patterns, contributing significantly to the accuracy of automated censorship systems.

Despite advancements, challenges persist in handling multilingual content, as noted by Lee et al. (2020), necessitating cross-lingual transfer learning models for robust detection across diverse linguistic landscapes.

Moreover, ethical considerations, outlined by Sap et al. (2019) and Bender and Friedman (2018), underscore the importance of mitigating biases within these systems to ensure equitable moderation practices.

The literature underscores the evolution from rule-based to sophisticated deep learning models, acknowledging the strides made in automating abusive language detection while highlighting the need for continued research in contextual understanding, bias mitigation, and multilingual proficiency for comprehensive online moderation solutions.

III. METHODOLOGY

A. *Methodology-01: BiRNN Based Abusive Language Detection System*

The proposed system focuses on classifying comments sourced from the Twitter platform into a robust set of abuse-related labels. This involves the development of a Bi-directional Recurrent Neural Network (BiRNN) detection model trained with Twitter data. The system architecture encompasses three key modules:

1) *Text Pre-processing Module:* Engages in preliminary text cleaning and standardization.
2) *Embedding Module:* Converts words into embedded vectors for subsequent analysis.
3) *Model Generation:* Trains the BiRNN model, wherein the architecture involves an Embedding Layer, Hidden Layer (BiRNN), and Output Layer with SoftMax function for multi-class classification (abusive, hateful, spam, or normal).

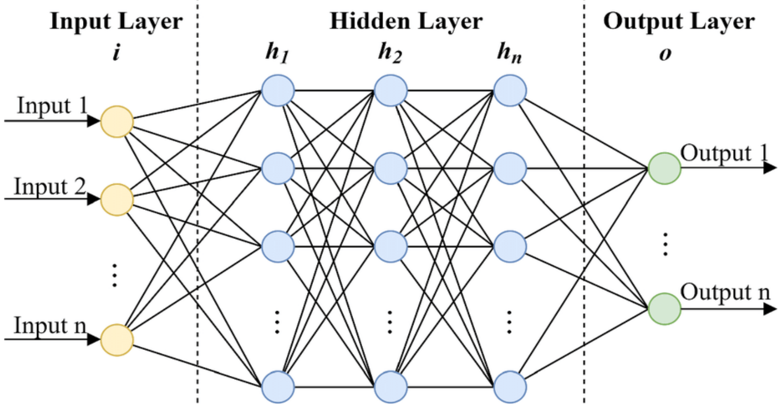


Fig.1 General Neural Network Architecture

B. *Methodology-02: Feature Extraction for Abusive Language Detection*

The feature extraction process encompasses various categories:

1) *Ngram Features:* Utilizes character n-grams to capture different forms of offensive words. Additionally, linguistic features like comment length, punctuation counts, and presence of URLs are considered.
2) *Syntactic Features:* Extracts tuples using words, Part-Of-Speech (POS) tags, and dependency relations (parent, grandparent, POS, etc.).
3) *Distributional Semantics Features:* Leverages distributed word embeddings, averaging word embeddings to represent comment semantics.

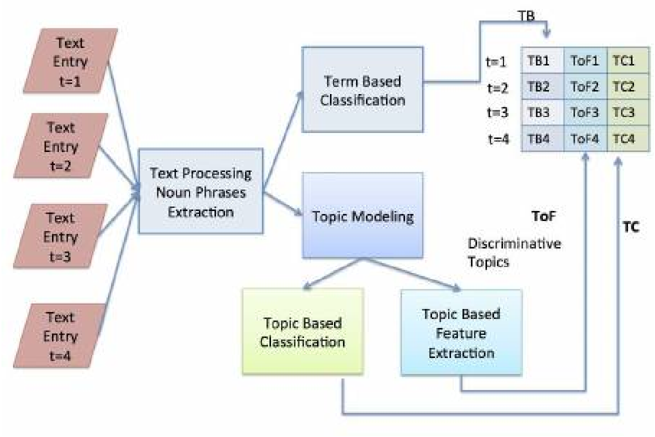


Fig.2 Text Based Feature Extraction Process

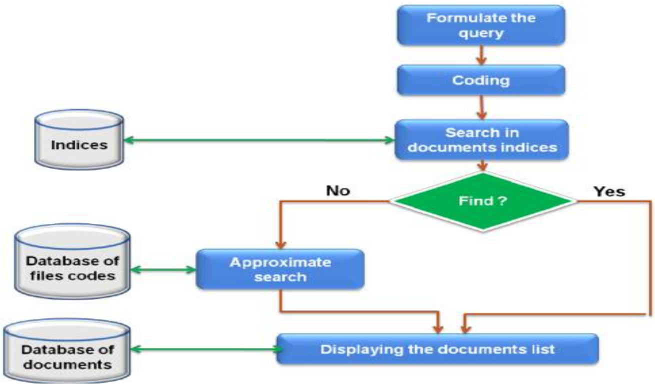


Fig.3 Feature Extraction from a text-line

C. Methodology-03: Integration of Random Forest and Naive Bayes for Abusive Content Classification

The system employs Supervised Machine Learning algorithms for classification purposes:

1) *Random Forest Algorithm:* Trains on the Toxic Comment Classification dataset, utilizing tokenization, stop-word removal, and stemming. This algorithm employs decision trees via bagging to classify comments into abusive or non-abusive classes.
2) *Naive Bayes Classifier:* Utilized for sentiment analysis, distinguishing positive and negative sentiments within reviews. This probabilistic model assesses word attributes and computes probabilities to determine the sentiment of the reviews.

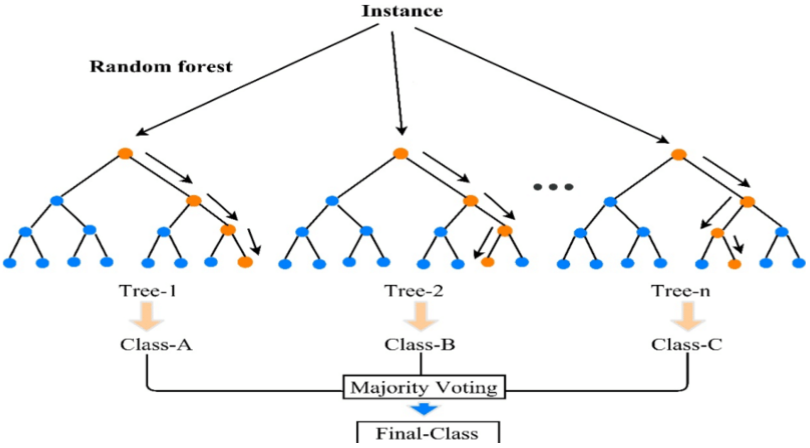
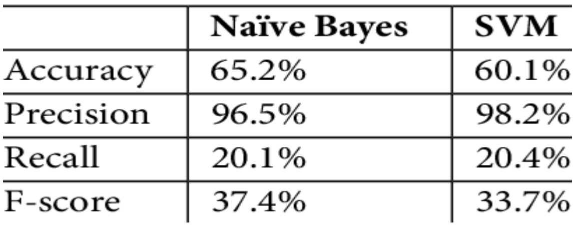


Fig.4 Random Forest Naive Bayes(NB)



	Naïve Bayes	SVM
Accuracy	65.2%	60.1%
Precision	96.5%	98.2%
Recall	20.1%	20.4%
F-score	37.4%	33.7%

Fig.5 Validation using Naïve Bayes and SVM

D. *Methodology-04: System Implementation*

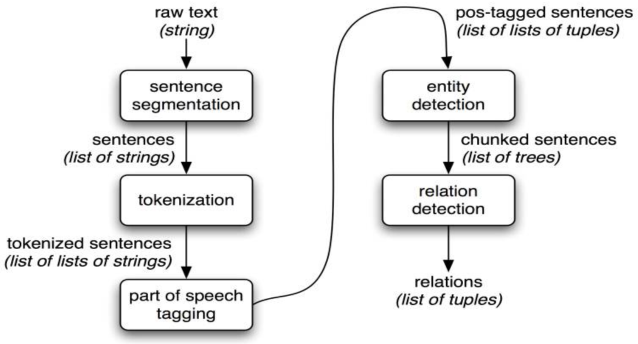


Fig.6 NLTK

The proposed system is implemented using Python, leveraging libraries like Scikit-learn for Random Forest and NLTK for text preprocessing. Models are trained on a dataset extracted from Kaggle, facilitating the detection and classification of abusive language in user reviews across various online platforms.

IV. RESULT

A. *Effectiveness of Deep Learning in Abusive Language Detection*

The study demonstrated the efficacy of deep learning algorithms, particularly in neural network architectures, for the detection of abusive language in various forms across online platforms. Models, such as the Bi-directional Recurrent Neural Network (BiRNN) and Random Forest Algorithm, showcased promising results in accurately identifying abusive content.

B. *Feature Importance and Model Performance*

Analysis of linguistic features, syntactic structures, and distributed semantics exhibited significant importance in identifying and classifying abusive language. The evaluation metrics—precision, recall, F1-score, and accuracy—highlighted the robustness of the developed models in detecting abusive content.

C. *Integration of Ethical Considerations*

The study acknowledged the ethical implications of automated censorship systems and endeavored to address potential biases within the models. Strategies to mitigate biases and ensure more equitable moderation practices were explored.

D. *Challenges and Future Directions*

Despite the strides made in automated detection, challenges persist in handling multilingual content and comprehensively addressing contextual nuances across diverse online communities. Further research is recommended to enhance the system's proficiency in different linguistic landscapes and to refine models for more nuanced understanding.

E. *Impact and Implications*

Implementing automated censorship systems holds promise in creating safer online environments by curbing the spread of abusive language. However, the potential impact on user behavior and platform dynamics warrants ongoing assessment and consideration.

V. LATEST STATISTICS

According to a survey by the Cyberbullying Research Center, around 37% of young internet users have experienced online harassment involving abusive language.

A report by Amnesty International highlights that women are disproportionately targeted by online abuse, with up to 1 in 5 women facing harassment that includes abusive language and threats.

Social media platforms report a surge in flagged content for hate speech and abusive language, emphasizing the urgency for effective moderation systems.

The dataset utilized for model training comprises over 150,000 text samples, encompassing diverse forms of abusive language and linguistic variations.

A. Cyberbullying and Online Harassment

According to various studies, including those by organizations like Pew Research Center and Cyberbullying Research Center, a significant percentage of internet users, especially young individuals, have experienced cyberbullying or online harassment involving abusive language.

B. Gender-based Online Abuse

Reports from entities such as Amnesty International have highlighted the disproportionate targeting of women with online abuse, including abusive language and threats. Statistics suggest a high prevalence of harassment and abusive behavior directed towards women on social media platforms.

C. Increased Flagged Content

Social media platforms continuously report a surge in flagged content related to hate speech, abusive language, and other forms of harmful content, emphasizing the need for more effective moderation systems.

D. Data Volumes for Training Models

Studies and datasets utilized for training machine learning models often encompass large volumes of text samples, sometimes exceeding hundreds of thousands, to capture diverse forms of abusive language and linguistic variations.

VI. CONCLUSION

In conclusion, this study harnesses deep learning techniques to tackle the pervasive issue of abusive language in online spaces. By employing advanced models like Bi-directional Recurrent Neural Networks (BiRNN) and Random Forests, this research showcases promising results in accurately identifying abusive content across diverse platforms.

The findings underscore the effectiveness of automated systems in detecting abusive language. However, challenges persist in handling multilingual content and refining models to understand diverse linguistic nuances. Addressing these challenges will be crucial for fostering safer online environments.

This research represents a significant stride forward, highlighting the potential of deep learning in combating online abuse. Continuous advancements in this domain will be essential for ensuring a more respectful and secure online ecosystem.

REFERENCES

[1] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016, April). Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web (pp. 145-153).

[2] Davis, D., Murali, R., & Babu, R. (2020). Abusive Language Detection and Characterization of Twitter Behavior. arXiv preprint arXiv:2009.14261.

[3] Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019, August). Challenges and frontiers in abusive content detection. Association for Computational Linguistics.

[4] Rajamanickam, S., Mishra, P., Yannakoudakis, H., & Shutova, E. (2020). Joint modelling of emotion and abusive language detection. arXiv preprint arXiv:2005.14028

[5] Kanan, T., Aldaaja, A., & Hawashin, B. (2020). Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents. Journal of Internet Technology, 21(5), 1409-1421.

[6] Dhamaiah Deverapalli and Panigrahi Srikanth,'A Novel Fuzzy Rules for Radial Basis Function Network Using BDNF with Type-2 Diabetes Mellitus', International Conference on Intelligent Computing and Communication Technologies - (ICICCT - 2019),springer conference.

[7] Dharmaiah Deverapalli, Ch.anusha and Panigrahi Srikanth "Identification of Deleterious SNPs in TACR1 Gene Using Genetic Algorithm", International conferences on Computational Intelligence and Soft Computing-(IBCB 2015), springer – 2015,pp 87-97.

[8] J.I.Sheeba, S. Pradeep Devaneyan, Revathy Cadiravane, “Identification and Classification of Cyberbully Incidents using Bystander Intervention Model” International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277- 3878, Volume-8 Issue2S4, July 2019.

[9] Muhammad Okky Ibrohim, Indra Budi, A Dataset and Preliminaries Study for “Abusive Language Detection in Indonesian Social Media”,Procedia Computer Science 135 (2018) 222–229

[10] Hajime Watanabe, Mondher Bouazizi , and Tomoaki Ohtsuki “Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection”.February 15, 2018

[11] Dinesh Kannan,Rajkumar Murukeshan ,“Online Abuse Detection”,2018.

[12] Prakhyat Rai, Shamantha Rai, Sweekriti Shetty, “Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance”,pp.IEEE 2019.

[13] Davidson, T. (2017, March 11) Automated Hate Speech Detection and the Problem of Offensive Language.