Guardian Voice – A Speech Safeguard System for Enabling Ethical Social Interactions

[1,2,3,4] Harsha S[1], Harinisri R[2], Charumathi U[3], Lavanya K[4], Janani.G[5]
[1,2,3,4]Students, [5]Assistant Professor, [1,2,3,4,5]Department of Information Technology,
RMD Engineering College, Tiruvallur, Tamil Nadu

Abstract: Work on voice sciences over recent decades has led to a proliferation of acoustic parameters that are used quite selectively and are not always extracted in a similar fashion. With many independent teams working in different research areas, shared standards become an essential safeguard to ensure compliance with state-of-the-art methods allowing appropriate comparison of results across studies and potential integration and combination of extraction and recognition systems. Guardian Voice presents an innovative solution aimed at monitoring and safeguarding social media audio interactions. Utilizing Artificial Intelligence (AI) and Natural Language Processing (NLP), this system operates in real-time to oversee conversations within social media platforms. Its primary objective is to detect and prevent the dissemination of unethical or harmful speech, ensuring a secure and respectful environment for users. Guardian Voice's proactive moderation capabilities address inappropriate content swiftly, promoting ethical dialogues and enhancing the overall user experience. Additionally, the system continually refines its algorithms to adapt to evolving speech patterns, ensuring effective and up-to-date content moderation. Ultimately, Guardian Voice strives to foster a positive and safe atmosphere within social media audio interactions. features that make this technology a game-changer for the global supply chain.

Keywords:
Abusive language; Artificial Intelligence; cyber bullying; Deep learning; hate speech detection; human computer interaction; machine learning; Natural Language Processing; social media; sound recognition; speech recognition; Verbal abuse;

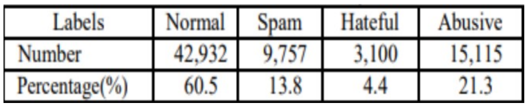
I. INTRODUCTION:

Online anonymity provides freedom of speech to many people and lets them speak their opinions in public. However, anonymous speech also has a negative impact on society and individuals. With anonymity safeguards, individuals easily express hatred against others based on their superficial characteristics such as gender, sexual orientation, and age. In this digital era, social media and social life has become a most important aspect of living a life. It is not only a source for information but also a medium for entertainment. It also allows people to express their opinions and their feelings about anything at any time. On onr such as Monte Carlo Simulation and Markov Chain Monte Carlo, address uncertainty and randomness within systems. Through an in-depth exploration of these optimization methodologies, this paper aims to demonstrate their applicability and effectiveness in various engineering fields, showcasing how they contribute to the overall feelings about anything at any time. On one hand, it can provide a medium for constructive criticism and for spreading positivity. On the other hand, the freedom of expression has caused some serious issue – Cyberbullying Cyberbullying is bullying that takes place over digital devices like cell phones, computers, tablets. Cyberbullying can occur through SMS, Text, and apps, or online in social media, forums, or gaming where people can view, participate in, or share content. This is due to anonymity given to the users and lack of sufficient regulations.

Automatically detecting the abusive words from the social media comment is a challenging task considering the user privacy regulations implemented in the social media algorithms. There is always a real chance that someone will be harassed online. Guardian Voice signifies a pioneering advancement in social media content moderation, specifically focusing on audio interactions. In response to the growing influence of audio-based communication on social platforms, Guardian Voice emerges as a solution harnessing Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies. Its fundamental purpose is to monitor, analyze, and ensure the ethicality of conversations occurring within these platforms. By leveraging real-time monitoring capabilities, Guardian Voice proactively identifies and prevents the dissemination of unethical or harmful speech, thereby fostering a secure and respectful environment for users. This introduction marks a pivotal step towards elevating content moderation in the realm of social media audio interactions, emphasizing the significance of ethical dialogues and user safety.

II. DATASETS

The lack of a standard dataset for hate speech detection makes it difficult to compare procedures and results using various data and comments. The datasets were developed for various goals, therefore they have distinct properties and show various forms of hate speech. Creating datasets for this activity takes time because the number of hateful instances in social networks is small, but a dataset must include a significant number of such cases. A number of datasets are also hidden to the general public. This could be due to concerns about privacy or the nature of the datasets, such as rudeness and inappropriate language. As previously noted goal of this suggested system is to create a classifier that uses BiRNN to classify text for a specific user comment. Table 1 displays the distribution of the 100k tweets in this dataset.



Labels	Normal	Spam	Hateful	Abusive
Number	42,932	9,757	3,100	15,115
Percentage(%)	60.5	13.8	4.4	21.3

Fig. 2.1 Example dataset

III. LITERATURE SURVEY

In paper [1] Nobata, Chikashi, et al. "Abusive language detection in online user content." Proceedings of the 25th international conference on world wide web. 2016.In addition to addressing the aforementioned gap in the area, algorithms presented in this research seek to create a cutting-edge way for identifying offensive language in user comments. The contributions made in this study are as follows:To outperform a deep learning system, this research uses supervised classification methodology with NLP features. utilises and modifies a number of the previous art features in an effort to compare their performance with the same data set. Add features from distributional semantics techniques to the feature list as well.Making a fresh data set of a few thousand user comments gathered from various domains public. This set comprises three judgments per remark and, for those considered abusive, a more detailed assessment of each comment's nature.

In paper [2]Davis, Dincy, Reena Murali, and Remesh Babu. "Abusive Language Detection and Characterization of Twitter Behavior." arXiv preprint arXiv:2009.14261 (2020).The main goal is to concentrate on numerous abusive behaviours on Twitter and determine whether or not a communication is abusive. The suggested BiRNN is a better deep learning model for automatically detecting abusive speech, according to results of comparisons between Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) approaches for various abusive behaviours in social media.The suggested method for detecting abusive language was assessed using two measures, namely accuracy and F1-measure. In the future, the effectiveness of the proposed system will be assessed using a variety of domain datasets, including those from Facebook, Wikipedia, Twitter, and other online communities, in order to generalise user behaviour.

In paper [3] Vidgen, Bertie, et al. "Challenges and frontiers in abusive content detection." Association for Computational Linguistics,2019.This paper discusses about the issues constrain, the performance, efficiency and generalizability of abusive content detection systems. Abusive content detection is a pressing social challenge for which computational methods can have a hugely positive impact. In this paper methods deals with critical insights into the challenges and frontiers facing the use of computational methods to detect abusive content. They differ from most previous research by taking an interdisciplinary approach, routed in both the computational and social sciences.

In paper [4] Rajamanickam, Santhosh, et al. "Joint modelling of emotion and abusive language detection." arXiv preprint arXiv:2005.14028 (2020).Aiming to tackle this problem, the natural language processing (NLP) community has experimented with a range of techniques for abuse detection. While achieving substantial success,these methods have so far only focused on modelling the linguistic properties The aim of our work is to investigate the relationship between emotion and abuse detection, which is likely to be independent of the biases that may exist in the annotations. They proposed a new approach to abuse detection, which takes advantage of the affective features to gain auxiliary knowledge through an MTL framework

In paper [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.This paper mainly focused on detecting these phenomena on English text, few works studied this phenomenon on Arabic. To evaluate the performance of the classifiers, we use Recall, Precision, and F1-

Measure. Future scope: it has provided a comprehensive comparison that would aid future research works in this direction

IV. MACHINE LEARNING BASED SPEECH SAFEGUARD SYSTEM

Guardian Voice signifies a pioneering advancement in social media content moderation, specifically focusing on audio interactions. In response to the growing influence of audio-based communication on social platforms, Guardian Voice emerges as a solution harnessing Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies. Its fundamental purpose is to monitor, analyze, and ensure the ethicality of conversations occurring within these platforms. By leveraging real-time monitoring capabilities, Guardian Voice proactively identifies and prevents the dissemination of unethical or harmful speech, thereby fostering a secure and respectful environment for users. This introduction marks a pivotal step towards elevating content moderation in the realm of social media audio interactions, emphasizing the significance of ethical dialogues and user safety.

Guardian Voice is a groundbreaking system designed to uphold ethical standards in social media audio interactions. Its scope encompasses real-time monitoring and analysis of conversations across various platforms. By employing advanced algorithms, it identifies and addresses potentially harmful or inappropriate content. This system aims to create a safer and more responsible online environment by promoting respectful communication. Guardian Voice offers comprehensive safeguards against cyberbullying, hate speech, and other forms of misconduct. Through proactive intervention, it mitigates the spread of harmful content and fosters constructive dialogue. The technology behind Guardian Voice enables seamless integration with existing social media platforms, ensuring widespread adoption and impact. Its robust features include sentiment analysis, voice recognition, and context-aware filtering to accurately assess conversations. Guardian Voice operates with a commitment to user privacy and data security, maintaining strict compliance with relevant regulations. Ultimately, it empowers users to engage in ethical and meaningful interactions while preserving the integrity of social media communities.

Guardian Voice is designed to monitor social media audio interactions using AI and NLP in real-time. Its primary objective is to detect and prevent unethical speech, ensuring a safe environment. The system aims to proactively moderate conversations, refine algorithms for better adaptability, and foster ethical dialogues within social media platforms.

a) **Increased security** : By using machine learning technology, security can be enhanced by training the model to find the abusive words and beep them.

b) **Enhancing user experience**: The guardian voice will enhance the user experience by identifying the abusive comments and hiding it so that the user will have a seamless experience.

c) **Parental control:** By using Guardian voice parental control can be achieved so that the parents need not to worry about their child’s mental health.

d) **Real time monitoring:** Activities are monitored closely and reported. Hence , any offensive or criminal use will be predicted earlier.

e) **Increased collaboration :** Without constant fear of getting abusive hatred comments, the social media influencers and the day to day users can be benefited by collaborating in the social media platform.

f) **Eliminate Cyberbullying :** Cyberbullying is one of the most uncontrollable drawback of using social media. With the help of guardian voice

V. WORKING OF SPEECH RECOGNITION

Speech recognition is a sophisticated technology that takes spoken language and converts it into written text. This process involves several important steps to ensure accuracy — from capturing audio, to language modeling.

a) **Capturing audio**: The first step is capturing the speech signal. The audio is recorded at a high rate to make sure all the details are captured, and then it's cleaned up to remove any background noise or interference. This helps to improve the quality of the speech signal. The noise can be removed from the speech signal through various filtering techniques, such as spectral subtraction, Wiener filtering, or Kalman filtering. These techniques estimate the presence of noise in the signal and subtract it, resulting in a clearer speech signal

b) **Identifying key Features and characteristics**: Next, the system identifies the key characteristics of the speech signal, called features. These features are used to differentiate between different speech sounds, such as Mel-Frequency Cepstral Coefficients (MFCCs) and pitch and energy features.
c) **Acoustic modeling**: The system is then trained on a large corpus of speech data to build an acoustic model. This model maps the speech features to their corresponding phonemes or sub-word units, allowing the system to identify the words being spoken.
d) **Language modeling**: In order to construct a language model, the system is additionally trained on a vast corpus of text data. This model represents the likelihood of word sequences in the language and assists the system in making informed judgments about the most likely words to be spoken in a particular context.
e) **Decoding:** Finally, the system uses acoustic and language models to transcribe the speech signal into text. It searches for the sequence of words with the highest probability given the models and outputs the text.

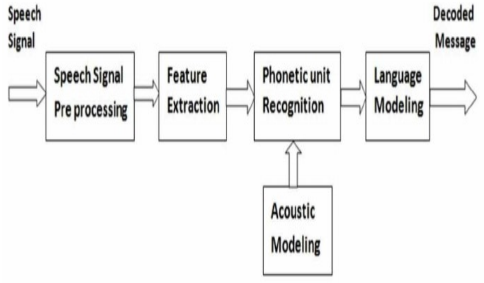


Fig. 4.1 Block diagram of speech recognition

VI. CHALLENGES AND OPPORTUNITIES

The literature review and analysis presented in previous sections provide insights into the current state of the art of abusive language studies in Indonesian. Based on these analyses, we have observed several challenges in this task, which are summarized as follows:

a) **Limited Availability of Language Resources**: The adopted approach for dealing with the task of abusive language detection in Indonesian is currently limited and lags behind studies in other, more resource-rich languages. Traditional models are the most popular approach for addressing this problem in Indonesia, while in other languages, more recent transformer-based models are commonly used to achieve state-of-the-art results. We believe that this discrepancy is likely related to the limited availability of language resources, including language corpora and language models.

b) **Limited Exploration of Abusive Phenomena**: Based on the abusive phenomena covered in the available datasets for abusive language detection studies, perceive that the explored abusive phenomena in Indonesian is still very limited. Studies in Indonesian have mostly focused on the detection of hate and abusive speech. Meanwhile, similar studies in other languages have been conducted with a broader coveage of abusive phenomena, which can include sexism, racism, misogyny, Islamophobia, and more. Some of these studies have also proposed finer-grained labels to capture more specific abusive phenomena, which is usually beneficial for differentiating the treatment for handling each phenomenon.

c) **Low Awareness of Reproducibility Aspect:** Based on our review, we also notice that most of the published research in Indonesian abusive language studies do not make their code and datasets publicly available. This issue makes it difficult for other researchers to reproduce the results of previous works, which is important for better analysis of their own studies. Furthermore, reproducibility is an important aspect for maintaining continuity in research, specifically in the area of abusive language research.

d) **Limited Approach for Annotation Procedure:** We observe that most studies used manual expert annotation procedures to label abusive language datasets. This approach is proven to be reliable for obtaining a high-quality dataset when the subjectivity of the annotation task is high. However, this approach is usually not feasible for annotating a large number of data, as the annotation task becomes more labor intensive and time-consuming. Sometimes, alternative annotation approaches such as crowdsourcing scenarios can provide a wider perspective, with a diverse demographic of annotators who have different backgrounds and views to evaluate the abusive instances.

e) **The Problem of Code-Mixed Languages:** Geographically, Indonesia consists of several regions, each with its own local languages. According to recent reports, there are 718 local languages used by different regions and tribes in Indonesia. Indonesians tend to use a mix of their own local language and Bahasa to communicate on social media platforms, such as Twitter. Related to this issue, we conducted a random check on some publicly available datasets. We found a lot of code-mixed instances on the checked datasets [28], [24], which are mostly written in a mixture of Indonesian and Javanese. As in other languages and other NLP tasks, the issue of code-mixing is still a prominent challenge that needs to be tackled. Based on these challenges, we also point out several opportunities for future studies in this research direction, which are summarized below.

f) **Building Novel Language Resources in Indonesian:** Our NLP research community should also focus on studying and developing language resources in Indonesian. These resources could include novel corpora for diverse tasks or recent language model technologies. The availability of more language resources could provide more opportunities for researchers in abusive language studies to explore more approaches to better detect abusive language in Indonesian.

g) **Expanding the Study Exploration into Other Abusive Phenomena:** As mentioned in the challenges section, abusive language studies in Indonesian are still focused on a few phenomena, including hate speech and abusiveness. Based on our investigation, there are several abusive phenomena specific to Indonesia that could potentially become a focus for exploration, including islamophobia and political hate speech. There are also other more general phenomena which have been studied in other languages, such as sexism, racism, xenophobia, homophobia, and more. A broader exploration into other abusive phenomena could open more opportunities for research collaboration between NLP researchers and researchers from other communities such as the study of humanity, psychology, gender studies, and social science.

h) **Exploring Other Annotation Approach to Build Abusive Language Datasets**: Most of the available abusive language datasets in Indonesian were built using expert annotation approaches. For example, crowdsourcing could be a worth-considering option to be implemented for annotating abusive language datasets. Because crowdsourcing approach has the advantage of bringing in a diverse set of annotators with different background identities, which can help to reduce bias in the dataset, which is also an important issue in this study. In addition, crowdsourcing can be particularly useful when the dataset is large and complex, and would be too time-consuming for a single person to finish.

i) **Tackling the Problem of Code-Mixed Data:** Codemixing is becoming a prominent challenge in various NLP tasks in recent years. This problem may be due to the current technology and platforms which have a multilingual environment. Similarly, Indonesians also tend to use a mix of their local languages and Bahasa Indonesia to communicate with others both in real life and on social media channels. Dealing with language-shift in code-mixed data is a challenging task. Specifically in abusive language studies, several transfer learning approaches could be applied in this task.

VII. SYSTEM ARCHITECTURE

Guardian Voice presents an innovative solution aimed at monitoring and safeguarding social media audio interactions. Utilizing Artificial Intelligence (AI) and Natural Language Processing (NLP). This system operates in real-time to oversee conversations within social media platforms. Its primary objective is to detect and prevent the dissemination of unethical or harmful speech, ensuring a secure and respectful environment for users.

The system architecture of guardian voice is depicted in the figure 5.1.

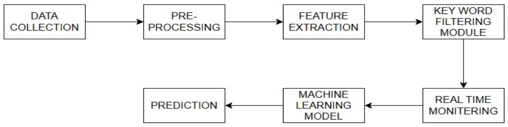


Fig. 5.1 System architecture of guardian voice

a) **Data Collection:** Gather a dataset of records that includes information about voice and their words , and whether they have had a previous results.This dataset should be large and diverse to train a robust machine learning model.

b) **Data Preprocessing :** Clean and preprocess the data. This involves handling missing values, data normalization, and converting categorical data into numerical form. Ensures data privacy and compliance with relevant regulations when dealing with sensitive speech data.

c) **Feature Selection and Engineering:** Identify relevant features (attributes) that can influence the speech prediction. These may include the specific attributes by its nature of the commodity and its characters and its nature by its following. Perform feature engineering to create new features or transform existing ones if needed.

d) **Data Splitting:** Split the dataset into training and testing sets. The training set is used to train the machine learning model, and the testing set is used to evaluate its performance.

e) **Machine Learning Model Selection:** Choose an appropriate machine learning algorithm for classification. Common choices include the values by using the keyword filter, speech recognition, GTTs to identify the values to find it.

f) **Model Training:** Train the selected machine learning model on the training data. The model learns to predict whether a voice and its attributes based on the provided features.

g) **Model Evaluation:** Evaluate the model's performance on the testing set using appropriate metrics, such as accuracy, precision, recall, to assess its predictive accuracy and reliability.

VIII. SYSTEM IMPLEMENTATION

The implementation of guardian voice is explained in the following sections :

a) **Exploratory data analysis:** During this step we performed some descriptive analysis and determined the target variable. We also explored how many classes were in the target and a selection of other possibly problematic (high cardinality) variables. I also visualized the target variable in a histogram which is a good technique for understanding the distribution of the data to assist in parameter tuning

b) **Data Cleaning:** Data cleaning is a crucial step in the machine learning (ML) pipeline, as it involves identifying and removing any missing, duplicate, or irrelevant data. The goal of data cleaning is to ensure that the data is accurate, consistent, and free of errors, as incorrect or inconsistent data can negatively impact the performance of the ML model. Data cleaning, also known as data cleansing or data preprocessing, is a crucial step in the data science pipeline that involves identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data to improve its quality and usability. Data cleaning is essential because raw data is often noisy, incomplete, and inconsistent, which can negatively impact the accuracy and reliability of the insights derived from it. For decision-making, the integrity of the conclusions drawn heavily relies on the cleanliness of the underlying data. Without proper data cleaning, inaccuracies, outliers, missing values, and inconsistencies can compromise the validity of analytical results. Moreover, clean data facilitates more effective modeling and pattern recognition, as algorithms perform optimally when fed high-quality, error-free input**.**

c) **Preprocessing text and transistion :**We removed the target variable from the entire data set and transformed the categorical variable into a model matrix with one-hot encoding. This is sometimes the requirements for certain algorithms to process the data in a sparse matrix format. Other statistical software such as R, automates this step when generating models. I imputed the missing values in the data to 0. I scaled the continuous variables using min-max normalization which transforms values onto a scale from 0 to 1 to prevent variables on different scales heavily impacting the coefficients.

d) **Data partition and evaluation :** We removed the target variable from the entire data set and transformed the categorical variable into a model matrix with one-hot encoding. This is sometimes the requirements for certain algorithms to process the data in a sparse matrix format. Other statistical software such as R, automates this step when generating models. I imputed the missing values in the data to 0. I scaled the continuous variables using min-max normalization which transforms values onto a scale from 0 to 1

IX. CONCLUSION

Future enhancements for the provided code could involve several areas of development. Firstly, integrating more advanced speech recognition engines beyond Google's service could enhance accuracy and language support, broadening the application's usability. Expanding the list of "bad words" to cover a wider range of inappropriate language would improve censorship effectiveness. Implementing sophisticated natural language processing techniques could enable the algorithm to better understand context and tone, further refining its censorship capabilities. Additionally, introducing user authentication and moderation features would allow users to customize censorship preferences according to their individual needs. Integration with machine learning models could enable dynamic adjustment of censorship based on user feedback and evolving language trends, ensuring ongoing effectiveness.

Enhancing the user interface for improved accessibility and ease of use would enhance the overall user experience. Cloud-based speech recognition services could be leveraged for scalability and reliability. Real-time monitoring and alerting mechanisms could be implemented to detect and address potentially harmful speech content promptly. Compatibility with multiple audio input sources, such as uploaded files or streaming audio, would broaden the application's utility.

Lastly, optimization for performance and resource efficiency would ensure the application can support high volumes of concurrent users without compromising function.

In conclusion, the provided code demonstrates the integration of speech recognition, text processing, and text-to-speech conversion algorithms within a Flask web framework. It showcases the capability to convert spoken words into text, censor inappropriate language, and convert the processed text back into speech. Through these algorithms, the application offers real-time speech-to-text functionality with censorship features, enhancing user experience and promoting responsible communication. This implementation highlights the versatility and potential of speech recognition technology in developing interactive and accessible applications. However, further refinement and optimization may be necessary to improve accuracy and usability. Overall, the code serves as a foundation for creating speech-driven applications with enhanced functionality and user engagement.

REFERENCES:

[1] H. Rainie, J. Q. Anderson, and J. Albright, The future of free speech, trolls, anonymity and fake news online. Pew Research Center Washington, DC, 2017.

[2] B. Mathew, N. Kumar, P. Goyal, A. Mukherjee et al., “Analyzing the hate and counter speech accounts on Twitter,” arXiv preprint arXiv:1812.02712, 2018.

[3] E. W. Pamungkas, V. Basile, and V. Patti, “Misogyny detection in twitter: a multilingual and cross-domain study,” Information Processing & Management, vol. 57, no. 6, p. 102360, 2020.

[4] C.Nagarajan and M.Madheswaran - ‘Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter’ - Journal of ELECTRICAL ENGINEERING, Vol.63 (6), pp.365-372, Dec.2012.

[5] C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011.

[6] C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor & Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011.

[7] C.Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.

[8] Nagarajan C., Neelakrishnan G., Akila P., Fathima U., Sneha S. "Performance Analysis and Implementation of 89C51 Controller Based Solar Tracking System with Boost Converter" Journal of VLSI Design Tools &amp; Technology. 2022; 12(2): 34–41p.

[9] C. Nagarajan, G.Neelakrishnan, R. Janani, S.Maithili, G. Ramya "Investigation on Fault Analysis for Power Transformers Using Adaptive Differential Relay" Asian Journal of Electrical Science, Vol.11 No.1, pp: 1-8, 2022.

[10] G.Neelakrishnan, K.Anandhakumar, A.Prathap, S.Prakash "Performance Estimation of cascaded h-bridge MLI for HEV using SVPWM" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:750-756

[11] G.Neelakrishnan, S.N.Pruthika, P.T.Shalini, S.Soniya, "Perfromance Investigation of T-Source Inverter fed with Solar Cell" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:744-749

[12] C.Nagarajan and M.Madheswaran, "Analysis and Simulation of LCL Series Resonant Full Bridge Converter Using PWM Technique with Load Independent Operation" has been presented in ICTES'08, a IEEE / IET International Conference organized by M.G.R.University, Chennai.Vol.no.1, pp.190-195, Dec.2007

[13] M Suganthi, N Ramesh, "Treatment of water using natural zeolite as membrane filter", Journal of Environmental Protection and Ecology, Volume 23, Issue 2, pp: 520-530,2022

[14] M Suganthi, N Ramesh, CT Sivakumar, K Vidhya, "Physiochemical Analysis of Ground Water used for Domestic needs in the Area of Perundurai in Erode District", International Research Journal of Multidisciplinary Technovation, pp: 630-635, 2019

[15] V. Lingiardi, N. Carone, G. Semeraro, C. Musto, M. D'Amico, and S. Brena, "Mapping twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis," Behaviour & Information Technology, vol. 39, no. 7, pp. 711–721, 2020.

[16] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang, "How noisy social media text, how diffrnt social media sources?" in Proceedings of the Sixth International Joint Conference on Natural Language Processing. Nagoya, Japan: Asian Federation of Natural Language Processing, Oct. 2013, pp. 356–364. [Online]. Available: https://www.aclweb.org/anthology/I13-1041

[17] E. W. Pamungkas, V. Basile, and V. Patti, "Investigating the role of swear words in abusive language detection tasks," Language Resources and Evaluation, pp. 1–34, 2022.

N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4675–4684. [Online]. Available: https://aclanthology.org/D19-147E.

[18] W. Pamungkas, V. Basile, and V. Patti, "Towards multidomain and multilingual abusive language detection: a survey," Personal and Ubiquitous Computing, pp. 1–27, 2021.

[19] Y. Wirawanda and T. O. Wibowo, "Twitter: expressing hate speech behind tweeting," Profetik: Jurnal Komunikasi, vol. 11, no. 1, pp. 5–11, 2018.

[20] E. Fauziati, S. Suharyanto, A. S. Syahrullah, W. A. Pradana, and I. Nurcholis, "Hate language produced by indonesian figures in social media: From philosophical perspectives," WISDOM, vol. 3, no. 2, pp. 32–47, 2022.

[21] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the indonesian language: A dataset and preliminary study," in 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2017, pp. 233–238.

[22] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in Indonesian Twitter," in Proceedings of the Third Workshop on Abusive Language Online. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 46–57. [Online]. Available: https://www.aclweb.org/anthology/W19-3506

[23] A. R. Isnain, A. Sihabuddin, and Y. Suyanto, "Bidirectional long short term memory method and word2vec extraction approach for hate speech detection," IJCCS (Indonesian Journal of Computing and Cybernetics Systems), vol. 14, no. 2, pp. 169–178, 2020.

[24] M. A. Ibrahim, N. T. M. Sagala, S. Arifin, R. Nariswari, N. P. Murnaka, and P. W. Prasetyo, "Separating hate speech from abusive language on indonesian twitter," in 2022 International Conference on Data Science and Its Applications (ICoDSA). IEEE, 2022, pp. 187–191.

[25] M. O. Ibrohim and I. Budi, "A dataset and preliminaries study for abusive language detection in indonesian social media," Procedia Computer Science, vol. 135, pp. 222–229, 2018.

[26] D. R. K. Desrul and A. Romadhony, "Abusive language detection on indonesian online news comments," in 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). IEEE, 2019, pp. 320–325.