IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

DETECTING TAMIL TEXTUAL THREATS IN SOCIAL MEDIA USING ARTIFICIAL INTELLIGENCE

[1]ABISHEK A, [2]Thadeus Cruz Govindapillai
[1]Student, [2]Student
[1]Department of **ARTIFICIAL INTELLIGENCE**,
[1]Sri Krishna College of Technology, Coimbatore, India

Abstract: In recent decades, the proliferation of social networks has facilitated extensive communication among individuals globally through microblogging technologies. This has enabled the free exchange of ideas, feelings, and information on a massive scale and in real-time. However, this increased connectivity has also led to the proliferation of offensive language, including hate speech and cyberbullying, on social media platforms. Legislation in various countries seeks to address this issue, with measures such as Germany's Network Enforcement Act. Advanced technical approaches are being developed to assist social media platforms in enforcing these laws, including the detection of offensive language using artificial intelligence (AI) techniques. While offensive language detection has been explored in various languages, including English, German, Turkish, Hindi, Chinese, and Arabic, addressing this issue in Tamil poses its own challenges due to linguistic complexities and dialectical variations. This paper aims to explore the implemented AI approaches, performance measurement methodologies, and dataset details for offensive language detection in Tamil. By providing a comprehensive overview of existing research, this paper aims to guide future studies and facilitate a more unified approach to addressing the issue of offensive language in Tamil social media.

INTRODUCTION

Individuals are increasingly interconnected through social networks, enabling them to communicate on a massive scale in real-time. However, this connectivity has also led to the widespread proliferation of offensive language, including hate speech and cyberbullying. Legislation in various countries seeks to address this issue, with measures such as Germany's Network Enforcement Act. Advanced technical approaches, including AI, are being developed to assist in the detection of offensive language on social media platforms. While offensive language detection has been explored in various languages, addressing this issue in Tamil poses its own challenges due to linguistic complexities and dialectical variations. This paper aims to explore the implemented AI approaches, performance measurement methodologies, and dataset details for offensive language detection in Tamil.

BACKGROUND

2.1. Offensive Language

Defining offensive language is a nuanced task, heavily influenced by personal knowledge and cultural context. Communication is deemed offensive if it contains language intended to cause harm, pain, or anger. Hate speech and cyberbullying represent significant forms of offensive language that have seen a surge in prevalence on social media platforms.

2.1.1. Hate Speech

Hate speech refers to text aimed at harming individuals or inciting violence or social unrest. It encompasses the dissemination of racial, religious, extremist, or terrorist ideologies. Hate speech can manifest in various forms, including gender and religious hate speech.

2.1.2. Cyberbullying

Cyberbullying involves online attacks targeting specific individuals. Due to the rapid dissemination of online content, cyberbullying can have severe consequences, surpassing those of physical and verbal abuse. Cyberbullying encompasses various forms such as flame, masquerade, impersonation, harassment, outing, deceit, exclusion, and cyberstalking.

2.2. Tamil Language

Tamil is a Dravidian language primarily spoken in southern India, Sri Lanka, and Singapore, with significant diaspora populations worldwide. It boasts a rich literary tradition and is one of the oldest classical languages in the world. Tamil presents unique linguistic challenges for offensive language detection, given its complex morphology and dialectical variations.

2.2.1. Dialects of Tamil

Tamil dialects vary across regions and communities, with notable distinctions in vocabulary, pronunciation, and grammar. Major Tamil dialects include Brahmin Tamil, Kongu Tamil, Madras Tamil, and Sri Lankan Tamil.

2.2.2. Challenges Regarding Tamil

Detecting offensive language in Tamil poses several challenges, including the use of colloquialisms, slang, and regional variations. Additionally, Tamil script can be written in multiple scripts, including Tamil script, Tamil-Brahmi script, and Grantha script, further complicating language processing tasks.

2.3. Tamil Text Preprocessing

Preprocessing Tamil text involves cleaning and normalizing unstructured data obtained from social media platforms. This includes handling informal language, dialects, misspellings, and non-standard characters. Common preprocessing steps include removing elongations, diacritical markings, and extra characters, as well as tokenization, lemmatization, and stemming.

2.4. Feature Representation

Feature representation is crucial for identifying abusive speech in Tamil text. Various feature types and combinations have been explored in the context of offensive language detection. These include lexical features, syntactic features, semantic features, and stylistic features, among others. Finding the most effective feature representation approach is essential for accurate threat detection in Tamil social media content.

2.4.1. Language Models

Language models are textual representations based on probabilities, offering insights into the likelihood of word sequences occurring within a text. The Bag-of-Words (BoW) model is a commonly employed method for text representation, despite its disregard for word order. Another effective approach is N-grams, which capture sequences of N contiguous words or characters. Char N-gram, a variant, outperforms term N-grams in similar tasks due to its ability to capture character-level patterns. Additionally, the Term Frequency-Inverse Document Frequency (TF-IDF) technique is widely used in Natural Language Processing (NLP) tasks, although it may yield slightly inferior results compared to more advanced feature representations. Traditional methods like word and char N-grams typically exhibit lower performance compared to TF-IDF.

2.4.2. Word Embedding

Word embedding, also known as word vectors, provides numerical representations of words, aiding in language comprehension through mathematical methods. Word2Vec is a popular approach for generating word embeddings, offering two models: skip-gram and continuous bag of words (CBOW). AraVec is a pre-trained word embedding specifically designed for Arabic words, while FastText, developed by Facebook, leverages character n-grams to represent words. FastText's output vector for a word is computed as the sum of its constituent character n-grams.

2.5. Artificial Intelligence

Artificial Intelligence (AI) encompasses the development of intelligent systems capable of performing tasks and making decisions akin to humans. In today's digital landscape, vast amounts of data are generated from various sources, necessitating advanced processing techniques. AI is divided into several subfields, including machine learning (ML).

2.5.1. Machine Learning Algorithms

Machine learning is a pivotal component of AI, focusing on systems that learn and improve from data. Supervised learning relies on labeled training data, while unsupervised learning deals with unlabeled data to uncover patterns and insights. Deep Learning (DL), a subset of machine learning, employs multi-layer neural networks to learn intricate data representations. Various deep neural network architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformers, have been utilized to address text classification tasks, demonstrating remarkable performance across domains including NLP and text categorization.

For text classification, machine learning approaches can be broadly categorized into supervised learning, unsupervised learning, and deep learning.

a. Supervised learning: Supervised machine learning methods rely on labeled training data. These models are trained using examples where both input data and corresponding output labels are provided. Once trained, the model can predict the output labels for new, unseen input data based on the patterns learned during training.

b. Unsupervised learning: In unsupervised learning, the model is provided with input data that lacks explicit labels or categories. Instead, the goal is to discover hidden patterns or structures within the data. Unsupervised learning algorithms include techniques such as clustering, anomaly detection, and neural networks. These methods aim to group similar data points together or identify outliers without the need for labeled data.

c. Deep learning (DL): Deep learning is a subset of machine learning that utilizes neural networks with multiple layers to learn complex data representations. Deep learning has demonstrated remarkable performance across various domains, including voice

and handwriting recognition, computer vision, and natural language processing (NLP), including text classification. Various deep neural network architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformers, have been employed for text classification tasks. Transformers have gained popularity in NLP and text categorization due to their ability to capture long-range dependencies in text data using multi-head self-attention mechanisms, which replace traditional recurrent layers.

3.DATASETS USED IN PREVIOUS RESEARCH

3.1 General Offensive Classification:

Alakrot et al. [57]: Employed a dataset of Arabic YouTube comments for detecting offensive language. Pre-processing involved tokenization, filtering, normalization, and feature extraction using N-gram and word-level features. Support Vector Machine (SVM) achieved the best accuracy of 90.05%.

Mohaouchane et al. [58]: Applied deep learning models including CNN, Bi-LSTM, and combined CNN-LSTM on Arabic social media dataset. Utilized Arabic word embedding for representation and achieved high accuracy using CNN.

Abu Farha et al. [59]: Utilized the OSACT dataset to compare multitask deep learning models. Pre-processing involved normalization, tokenization, and feature extraction. Multitask learning classifiers, particularly MTL-S-N, achieved a macro F1 score of 90.4%.

Chowdhury et al. [47]: Developed a system for detecting abusive language using a multi-platform dataset. Pre-processing included tokenization and filtering out stop words, URLs, diacritics, and punctuation. SVM-based model achieved a performance of 84% using LOPO approach.

Keleg et al. [60]: Tested various classification models including AraBERT, MBERT, CNN, BiLSTM, and LR for offensive language detection using the OSACT dataset. Achieved high macro F1 score of 88% with AraBERT.

3.2 Hate Speech Classification:

Albadi et al. [61]: Constructed a hate speech classification model using Arabic lexicons and GRU-based RNN. Achieved an F1 score of 77% with the proposed model.

Haddad et al. [51]: Utilized the T-HSAB dataset for hate speech classification. Applied traditional machine learning classifiers such as SVM and NB, achieving an F1 score of 83.6% with NB.

Hassan et al. [4]: Used the OSACT4 dataset for hate speech classification. Employed convolutional neural networks and ensemble methods, achieving the top accuracy of 97.7%.

3. 3 Cyberbullying Classification:

Haidar et al. [56]: Developed a cyberbullying detection system using the AraBully-Tweets dataset. Pre-processing involved removing non-Arabic letters, user mentions, and single Arabic characters. Achieved an accuracy of 85.49% using SVM.

Haidar et al. [44]: Extended their research to incorporate deep learning models for cyberbullying detection. Utilized NB and SVM-based learning techniques, achieving an F1 score of 92.7% with SVM.

Haidar et al. [31]: Employed a FFNN model for cyberbullying detection, achieving a high accuracy of 92.53%.

These studies demonstrate the diverse methodologies and datasets utilized for detecting and classifying textual social threats in both Tamil and English languages, contributing to the development of effective AI-based solutions for addressing online safety concerns.

4.CONCLUSION:

Lately, there has been a surge of interest among researchers in detecting offensive language, both in Tamil and English, across social networks using artificial intelligence. This article delves into the suggested methods for identifying instances of offensive language in both languages, encompassing various forms such as hate speech and cyberbullying. The discussion includes an exploration of the techniques employed, performance metrics assessed, and characteristics of the datasets utilized, including dialect variations, annotation methods, and platform sources.

In reviewing the findings of these studies, it becomes evident that the study of offensive language detection, particularly in Tamil and English, is still in its nascent stages. Many of the techniques proposed have yet to be implemented into practical classification systems for text in these languages. Moreover, the availability of datasets specifically tailored for offensive language categorization in Tamil and English remains limited. Consequently, researchers face considerable challenges in this domain, including the need for more diverse and comprehensive datasets, complex pre-processing procedures, and a scarcity of publications addressing these issues.

REFERENCES

[1] S. Rathore, P. K. Sharma, V. Loia, Y. S. Jeong, and J. H. Park, "Social network security: Issues, challenges, threats, and solutions," Information Sciences, vol. 421, pp. 43–69, 2017, doi: 10.1016/j.ins.2017.08.063.

[2] K. Cortis and B. Davis, Over a Decade of Social Opinion Mining, vol. 2. Springer Netherlands, 2020.

[3] I. Science and H. H. Universit, Arabic Language Processing: From Theory to Practice, vol. 782, no. April. 2018.

[4] S. Hassan, Y. Samih, H. Mubarak, A. Abdelali, A. Rashed, and S. A. Chowdhury, "{ALT} Submission for {OSACT} Shared Task on Offensive Language Detection," Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, no. May, pp. 61–65, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.osact-1.9

[5] Samantha Kent, "German Hate Speech Detection on Twitter," Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), no. Konvens, pp. 120–124, 2018.

[6] F. Fortin, J. Delle Donne, and J. Knop, "The Use of Social Media in Intelligence and Its Impact on Police Work," in Policing in an Age of Reform, no. January, 2021, pp. 213–231.

[7] E. Hamdy, "Neural Models for Offensive Language Detection," arXiv, 2021.

[8] S. Mehra and M. Hasanuzzaman, "Detection of Offensive Language in Social Media Posts by Sidharth Mehra This thesis has been submitted in partial fulfillment for the," Thesis, Cork Institute of Technology, 2020.

[9] J. Risch, "Reader Comment Analysis on Online News Platforms," Thesis, Universität Potsdam, 2020.

[10] F. Husain and O. Uzuner, "A Survey of Offensive Language Detection for the Arabic Language," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 20, no. 1, pp. 1–44, 2021, doi: 10.1145/3421504.

[11] A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, "A systematic review of text classification research based on deep learning models in Arabic language," International Journal of Electrical and Computer Engineering, vol. 10, no. 6, pp. 6629–6643, 2020, doi: 10.11591/IJECE.V10I6.PP6629-6643.

[12] S. M. Abdou and A. M. Moussa, "Arabic Speech Recognition: Challenges and State of the Art," Computational Linguistics, Speech and Image Processing for Arabic Language, pp. 1–27, 2018, doi: 10.1142/9789813229396_0001.

[13] N. Y. Habash, Introduction to Arabic natural language processing, Synthesis Lectures on Human Language Technologies, vol. 3, no. 1. 2010.

[14] J. Risch, R. Ruff, and R. Krestel, "Explaining Offensive Language Detection," Journal for Language Technology and Computational Linguistics, vol. 34, no. 1, pp. 1–19, 2020.

[15] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, "Misinformation in Social Media: Definition, Manipulation, and Detection," ACM SIGKDD Explorations Newsletter, vol. 21, no. 2, pp. 80–90, 2019.

[16] R. Pradhan, A. Chaturvedi, A. Tripathi, and D. K. Sharma, "A review on offensive language detection," in Lecture Notes in Networks and Systems, vol. 94, no. January, 2020, pp. 433–439.

[17] G. Kovács, P. Alonso, and R. Saini, "Challenges of Hate Speech Detection in Social Media," SN Computer Science, vol. 2, no. 2, pp. 1–15, 2021, doi: 10.1007/s42979-021-00457-3.

[18] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: A survey on multilingual corpus," in Computer Science & Information Technology(CS & IT), Feb. 2019, no. March, pp. 83–100, doi: 10.5121/csit.2019.90208.

[19] O. Istaiteh, R. Al-Omoush, and S. Tedmori, "Racist and Sexist Hate Speech Detection: Literature Review," 2020 International Conference on Intelligent Data Science Technologies and Applications, IDSTA 2020, pp. 95–99, 2020, doi: 10.1109/IDSTA50958.2020.9264052.

[20] W. N. Hamiza Wan Ali, M. Mohd, and F. Fauzi, "Cyberbullying Detection: An Overview," Proceedings of the 2018 Cyber Resilience Conference, CRC 2018, no. November, pp. 1–3, 2019, doi: 10.1109/CR.2018.8626869.

[21] K. R. Talpur, S. S. Yuhaniz, N. N. B. A. Sjarif, B. Ali, and N. B. Kamaruddin, "Cyberbullying detection: Current trends and future directions," Journal of Theoretical and Applied Information Technology, vol. 98, no. 16, pp. 3197–3208, 2020.

[22] B. Haidar, M. Chamoun, and A. Serhrouchni, "A multilingual system for cyberbullying detection: Arabic content detection using machine learning," Advances in Science, Technology and Engineering Systems, vol. 2, no. 6, pp. 275–284, 2017, doi: 10.25046/aj020634.

[23] O. Oueslati, E. Cambria, M. Ben HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," Future Generation Computer Systems, vol. 112, pp. 408–430, 2020, doi: 10.1016/j.future.2020