iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61475

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Hate Speech Detection Using Deep Learning

Avishkar Gautam[1], Ayush Singh[2], Ayush Verma[3], Dr. Jaya Sinha[4]
Department of Computer Science and Engineering Galgotia's College of Engineering and Technology, Greater

Abstract: ***Hate speech is a form of verbal abuse that targets individuals depending on their race, religion, ethnicity, gender, sexual orientation, or other personal characteristics. It can be spoken, written, or displayed, and can be found in a variety of contexts, including online, in person, and in the media. Hate Speech has the potential to have a devastating impact on victims, causing emotional distress, social isolation, and even physical harm. Recognizing hate speech is a challenging task, especially in multilingual contexts. This is because hate speech can be expressed in many ways, and can be difficult to distinguish from other forms of speech. However, Artificial Intelligence (AI) has advanced recently, and have made it possible to develop effective hate speech detection systems. Hate speech detection is a challenging task, especially in multilingual contexts. This survey paper reviews the recent advances in hate speech detection using BERT and CNN models. We explored the various approaches that have been proposed, the challenges faced, and the paths that research will take in the future.***

I. INTRODUCTION

Throughout the previous years, the proliferation of online communication platforms has led to an exponential increase in the dissemination of hate speech, causing significant societal harm and posing substantial challenges to the maintenance of a healthy online environment. Hate speech, often defined as any form of communication that promotes hatred or incites violence directed at certain people or groups due to their characteristics such as race, ethnicity, religion, gender, or sexual orientation, has garnered widespread attention due to its potential to fuel discrimination, intolerance, and social unrest. With the advent of social media and the internet, hate speech has become increasingly prevalent, creating a pressing need for effective and efficient methods to detect and mitigate such harmful remark. Hate speech can be expressed in many different forms, including written text, spoken language, images, and videos.

In the last several years, there has been a growing awareness of the problem of hate speech, especially social networking platforms when used online like X, Facebook, Instagram and have evolved into breeding grounds for hate speech, as it enable users to communicate anonymously and without fear of repercussions.

The detection and classification of hate speech is a challenging task, especially in multilingual settings. Hate speech can be difficult to identify because it often relies on subtle cues, such as sarcasm, irony, and context. Additionally, hate speech can vary significantly from one language to another.

The ease and relatability of regional content have led to increased code-mixing of languages on social media, and this along with the increasing diversity of users present on such platforms is changing the way in which social media is used, and that too at an ever-increasing pace.

A standard model shall fail to identify profanity in content and distinguish its intensity, when applied to such blends of languages, given the numerosity of regional languages. The predominance of such code-mixed content on social media in the politically charged and volatile Indian Sub-Continent demands special attention for dealing with Abusive and Hate-Inducing content in Hinglish.

These platforms have also been shown to focus more on creative languages like Spanish, English, and so forth. This is due to the fact that the majority of the research being done on hate detection now covers common languages like English, Spanish, Turkish, etc. The provision of different local languages for comment writing on these platforms created additional obstacle in the identification process, making it difficult to identify hate remarks given the limited excellent work available for the low-resource Hindi language. Statistically, the Hindi language is the 4th primary spoken language globally, covering more than 330 million people. Furthermore, it is one of the two major official languages adopted by the Indian Constitution and covers around 46% population in India.

India is a land of diversity in terms of religion, living style, culture, language spoken, etc. Along with the various languages, each language has different dialects, which makes it more challenging to detect hate speech at ground level. Besides these wide varieties of languages, hate incidents are also increasing significantly every year. As per the data reported by the Government of India, the cases lodged under the hatred (Section 153A) are being increased yearly.

©IJRASET: All Rights are Reserved

Although Hindi is a resource constraint language, its colossaluser base serves as the salient source of information dissemination on social media. It shows how hate is disseminated using these platforms and makes society vulnerable to maintain peace and harmony. This work focuses on automatic hate speech detection forthe Hindi language. Often, hate speech is used to incite violence against a particular group. This type of speech is designed to stir uptrouble and ultimately does nothing to improve the lives of those targeted.

In contrast to Hate-inducing content, abusive texts contain some degree of profanity, but are only offensive in a vague sense. Though hurtful, such texts do not call for automatic intervention and steady removal, and can be dealt with lesser strictness, that too when reported. Example of hate speech for each class can be seen in TableI.

TABLE I. SAMPLE TWEETS

Text	Label
your tweet shows that indians are really k**fir.	Hate Speech
apne kaam se kaaam zyada bh*wa mat ban bjp ke ku**e	Hate Speech
#SacredGames kya bakwas series hai.	Non-Offensive
Galat insaan se toh sahi jaanvar hi accha hai! #CatPerson	Non-Offensive

II. LITERATURE SURVEY

This section outlines earlier research projects on hate speech identification:

A. Normalization of Lexical Variations

Depending on how words are spoken, Hindi written using the English script is referred to as Roman Hindi. Hindi in Roman script is not a common tongue Thus, many terms in its vocabulary do not have conventional spellings. Similar terms are spelled differently by various people. Because of this, it is challenging to convert the many lexical variations of Roman Hindi terminology to standard orthography, and not much study has been done in this field. An approach to feature-based clustering was developed by the authors in [1] It converts Roman Urdu words into phonetic equivalents and then groups together phonetically similar lexical variations of roman Hindi terms. The method was assessed using a manually marked gold standard dataset and outperformed the baseline in terms of F-score. Similarly, the authors of [2] suggested a technique that relies on a phonetic algorithm to make lexical variances in the Roman Urdu text more typical. The researchers from [2], [3], [4], [5], and [6] carried out a comparative analysis to evaluate how lexical variance is handled by normalization approaches in the text of multilingual social media postings, Including Roman Hindi, Dutch, Finnish, Arabic, Spanish, Bangla, Japanese, Chinese, and Polish. Among the several methods of normalizing is the Rule-based approach stemming and lemmatization, phonetic algorithms, machine learning algorithms, etc.

B. Lexical Rule-Based Approaches

Lexical rule-based techniques have been presented by researchers to identify offensive and phrases used in hate speech efficiently. The work of [7] provided an approach based on lexicons for detecting subjectivity in words in order to identify hate speech and it uses a process based on rules to create a lexicon of words associated with hatred. Next, a classifier is trained using the gathered information to identify hate speech, extracts from the lexicon were tested on the paper. Similarly, VADER, a rule-based method, was described by the authors of [8] for social media sentiment analysis of material. They created the Gold Standard Sentiment Lexicon and verified it They determined and examined the regulations on the customary use of textual grammatical and syntactic elements It contrasted the effectiveness of techniques based on lexical rules with the baseline models. The primary limitation of methods relying on lexical rules is their failure to accurately identify the context and word domain in a document. Furthermore, rule-based methods rely on lexicons. As a result, these techniques are not resistant to textual nuances and adversaries.

C. Traditional Machine Learning Approaches

Lexical techniques have been considered effective in the past for identifying perhaps offensive terms. However, further investigation showed that, aside from a few phrases that Human Coders recognized and the Hatebase lexicon specified, the lexical techniques produced inaccurate results in identifying hate speech.

On the other hand, machine learning algorithms demonstrated a considerable improvement in hate speech identification over lexical approaches [9]. The writers of [9] found that Linear Regression, Logistic Regression and SVM outperformed the other algorithms by a substantial margin in identifying hate speech in tweets. Furthermore, the research demonstrated that biases in the dataset and considering the fact that both hate speech contains sentences that overlap and misclassifications resulted from inflammatory tweets.

The study by [10] showed that every advanced model that has been suggested is susceptible to attackers. They demonstrated the limitations of adversarial training in resolving the problem and claimed that characteristics at the character level are more resilient to assault than ones at the word level. They recommended using character-level characteristics in logistic regression for the creation of a strong hate speech detection model. Multi-view SVM is a unique approach introduced by the authors of [11], It further offers enhanced interpretability and excelled the hate speech detection technologies from the past. Likewise, [12]'s study showed that the bigram characteristics, demonstrated the greatest overall accuracy (79%) for automated Hate Speech Detection when paired with the support vector machine technique. The SVM- Radial Basis Function (RBF) approach was proposed by the authors based on character level Using FastText to detect hate speech in social media texts with mixed Hindi and English codes. They showed how FastText's character-level features offer more data for categorization than word and document-level characteristics. The majority of attempts to identify hate speech have been done in English. English-language initiatives have been made to identify hate speech. The task of identifying hate speech in Roman Hindi has not gotten much attention. The writers of [13] used a range of machine learning techniques for identifying and proving hate speech in Roman Urdu because Logistic Regression fared better than other machine learning techniques using the deep learning method (CNN) in identifying tweets that are hostile and those that are neutral. Furthermore, they claimed that a word bag is a suitable feature extraction technique for identifying hate speech in Roman-Hindi. Conventional Machine Learning models work well when the dataset size is limited. Nevertheless, large datasets cause traditional machine learning models to perform worse.

D. Unsupervised Learning Approaches

Unsupervised learning techniques have been put out by a number of researchers to address the issue of hate speech identification. The Growing Hierarchical Self-Organizing Map (GHSOM), an unsupervised learning technique for social media cyberbullying detection, was suggested by the authors of [14].

The hand-crafted features acquired by this study were used to capture the syntactic and semantic characteristics of cyberbullies. The suggested method worked well for spotting cyberbullying on social media sites. Still, the method failed to recognize content that was satirical. The work is limited to the English language as well. The authors of [15] looked into a novel framework for identifying Facebook topics or issues that are commonly discussed and can lead to hate speech. Graphs, sentiment, and emotion analysis techniques were employed to group and examine postings on well-known Facebook sites.

As such, the suggested system has the ability to automatically identify websites that endorse hate speech in comment areas about delicate subjects. The findings showed that the accuracy of the suggested approach was 0.74. This work, however, is limited to hate speeches in English and employs slurs and the English language within the framework of American culture.

Hate speech is influenced by language and other demographic variables. There are differences between insults and insulting phrases in Roman Urdu and English. A method for unsupervised identifying German hate speech on Twitter was provided by the authors of [16]. They classified the terms into relevant categories, such as immigration, crime, and politics, using the k-mean clustering approach after using the skip-grams method to ascertain the words' context. Automatic hate speech detection was achieved through the means of a model for machine learning.

E. Deep Learning Approaches

While handling large datasets, Deep Learning surpasses traditional techniques. However, when working with restricted quantities of data, ordinary machine learning approaches are advised. To identify offensive information on social media networks[17]'s work suggested CNN with word2vec embedding and demonstrated how deep learning models might perform better than typical SVM classification results in cases when the training dataset was unbalanced. Oversampling can significantly increase the SVM's performance. This study helps researchers choose appropriate text categorization techniques for identifying harmful content, even when there is class imbalance in the training dataset. Similar to this, the writers of [18] carried out a number of investigations to find evidence of cyberbullying on several social media sites (SMPs). Four Deep Neural Network (DNN) models and three datasets—Formspring CNN, LSTM, Bidirectional LSTM (BiLSTM), with Attention were specifically implemented with 12,000 posts on Wikipedia, 16,000 posts on Twitter, and 100,000 posts on Twitter. Every DNN model made use of the same basic architecture as [19].

Since bullying constituted the minority class in the entirely imbalanced datasets, it was found that the models supported not bullying. It was also shown that oversampling significantly improved performance. Finally they found that on all three datasets, DNN models in conjunction with Transfer Learning outperformed state-of-the-art results.

By employing the BERT model for Abusive Language categorization, the authors of [20] showed that, when adjusted for the underlying issue, BERT outperforms other cutting-edge methods.For languages for which the BERT model has already been pre-trained, BERT can be adjusted. A multi-channel model with three BERT variants (MC-BERT) was proposed by the authors of [21] for the identification of hate speech. BERTs in multilingual, Chinese, and English. Furthermore, they translated training and test data into the languages needed by different BERT models in order to investigate the usage of translations as additional input. On three different datasets, they found that fine-tuning the Pre-trained BERT model produced state-of-the-art or equivalent performance. The lexical variances of Roman Urdu words make it very challenging to translate, and the language does not have a pre-trained BERT model. Additionally, dictionaries in both languages are needed for translation, because there isn't a Roman Urdu dictionary that has the terms' Standard Lexical form. In their study, [22] presented CNN-gram, a novel deep learning architecture for identifying abusive and hateful words in Roman Urdu, and evaluated its effectiveness on the RUHSOLD dataset against seven other baseline methods. When compared to the baselines, the suggested model demonstrated more robustness. The research by [23] showed that neural-based methods outperform classical machine learning techniques in the identification of hate speech. When it came to identifying hate speech, BI-LSTM with multiple embeddings performed the best among neural network architectures. The authors [24] evaluated the suggested model using two publicly accessible datasets and offered a transfer learning technique for identifying hate speech based on an already-existing pre-trained language model known as BERT. Subsequently, they devised a technique to mitigate the impact of bias in the training dataset while refining a pre-trained BERT-based model for the hate speech identification job. A pipeline for modifying the general-purpose RoBERT model to identify hate speech in Vietnamese was shown in the work by [25]. With the help of their suggested pipeline, the Vietnamese Hate Speech Detection (HSD) campaign achieved a new F1 score of 0.7221, a significant improvement in performance.

HateXplain, the first benchmark dataset of hate speech including several aspects of the subject, was introduced by [26]. They observed that even models that do remarkably well in classification do not do well on explainability criteria like model fidelity and plausibility, using state-of-the-art models at the moment. They also found that models with human reasoning used in training are better at reducing unintentional bias toward target populations. The study conducted by [18], [19] was carefully examined by [27], they came to the conclusion that, on certain datasets, the bulk of which are in English, supervised algorithms achieve almost flawless performance, according to the results produced by state-of-the-art systems. They looked at what seemed to be a gap between existing research and practical applications. They looked at the generalizability of the experimental methods used in earlier research [18], [19] to other datasets. Their findings showed a considerable dataset bias and methodological issues. Consequently, statements about current state-of-the-art performance have been significantly overstated. They concluded that sample issues and data overfitting account for most of the problems.

III. SUMMARY OF RELATED WORK

1) N. D. Gitari, Z. Zuping, H. Damien, and J. Long, proposed A lexicon-based approach for identifying hate speech. Proposed a lexicon-based method for detecting hate speech that creates a lexicon and finds subjectivity in phrases of hate- related words.
Limitations: The method is lexical rule-based and ignores the context and domain of terms inside a document.

2) C. Hutto and E. Gilbert, ‘‘VADER: A parsimonious rule- based model for sentiment analysis of social media text. Proposed VADER, a model built on rules of sentiment analysisof social media text. They created and validated the human- curated Gold Standard Sentiment Lexicon.
Limitations:

- The proposed method cannot identify if text messages constitute hate speech since it depends on a valence-aware vocabulary or lexicon and does not take into account the context.
- The suggested rule-based approach is not resistant to opponents or textual nuances because it is lexicon-based.

3) T. Davidson, D. Warmsley, M. Macy, and I. Weber, ‘‘Automated hate speech detection and the problem of offensivelanguage, Proposed a Logistic Regression with L regularization.Limitations:

- The work is limited to hate speech identification in the English language.
- Logistic regression may perform poorly on a huge dataset.

4) S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian,and O. Frieder, ‘‘Hate speech detection: Challenges and solutions. They devised the Mulli-view SVM approach, offering the added advantage of improved interpretability, which outperformed current systems.
Limitations: The analysis of hate speech in this study is restricted to English language and does not take user intent or context into account.

5) S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. Khan, and G. Mujtaba, ‘‘Automatic hate speech detection using machine learning: Bigram features combined with SVM performed best in automatic Hate Speech Detection with 79% accuracy.
Limitations:

- Text context cannot be captured by the proposed ML model, and thus cannot provide real-time predictions.
- As dataset sizes increase, classical models may not perform as expected.

6) H. Chen, S. McKeever, and S. J. Delany, demonstrated that CNNoutperforms SVM for abusive content identification when the training dataset is uneven. Showed Oversampling improves SVM performance beyond the deep learning model. Limitations: The SVM might not function properly on a balanced dataset if the dataset has millions of text messages in it.

7) S. Agrawal and A. Awekar, ‘‘Used Deep learning for detecting online attacks across multiple social media platforms, Showed that the Deep Learning model (Bit-STM) outperformed MachineLearning models for cyberbullying detection.
Limitations: They oversampled the whole dataset before the train-test split and before starting the cross-validation process in order to balance the classes, which causes overfitting.

8) I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, ‘‘ETHOS: A dataset to identify hate speech online. Presented a protocol for creating a suitable textual dataset called 'ETHOS', based on YouTube, Reddit comments. Demonstrated that performance of Neural-based approach (BiLSTM) is better thanClassical ML techniques.

9) Limitations: Since the HS detection dataset was limited to extracting English comments from social media, it was not multipurpose.

10) M. Mozafari, R. Farahbakhsh, and N. Crespi, ‘‘Hate speech detection and racial bias mitigation in social media based on BERT model. A loss-based fine-tuning technique is implemented to use newly reweighted training data to refine the BERT model that has been already trained.
Limitations: The work is limited to AAE/SAE languages; additional cross-domain datasets with various languages/dialects are not examined.

11) Q. H. Pham, V. Anh Nguyen, L. B. Doan, N. N. Tran, and T. M. Thanh Proposed a pipeline that substantially improves performance and obtained a new state-of-the-art level of performance in Vietnamese Hate Speech Detection HSD. Limitations: The identification of hate speech in Vietnamese is the only focus of this effort.

12) B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. This work introduced HateXplain, a new benchmark dataset for hate speech detection. Showed that the models, which utilize the human rationales for training, performed better in reducing unintended bias towards target communities.
Limitations:The work is only available in English. Hate speech that is multilingual was not taken into consideration. It is possible to use this method with different languages.

13) A. Rodriguez, C. Argueta, and Y.-L. Chen Proposed a novel unsupervised method based on graph analysis for identifying websites that may be disseminating hate speech.
Limitations: Systems with embeddings for Deep Learning can improve the suggested model.

IV. CONCLUSION

In the field of natural language processing (NLP), hate speech identification is an important topic, especially when it comes to online social media platforms. Given its rich linguistic history and extensive use, Hindi poses a particular challenge for the identification of hate speech because of its dialectal variances and cultural quirks. This survey research examined the most recent developments in the identification of hate speech.

The poll began by examining the traits of hate speech, emphasizing the difficulties caused by its innate ambiguity, cultural sensitivity, and code-switching. After that, the poll provided a thorough summary of the body of research on the identification of hate speech. It grouped the research according to the datasets used, methodological strategies used, and performance measures attained.

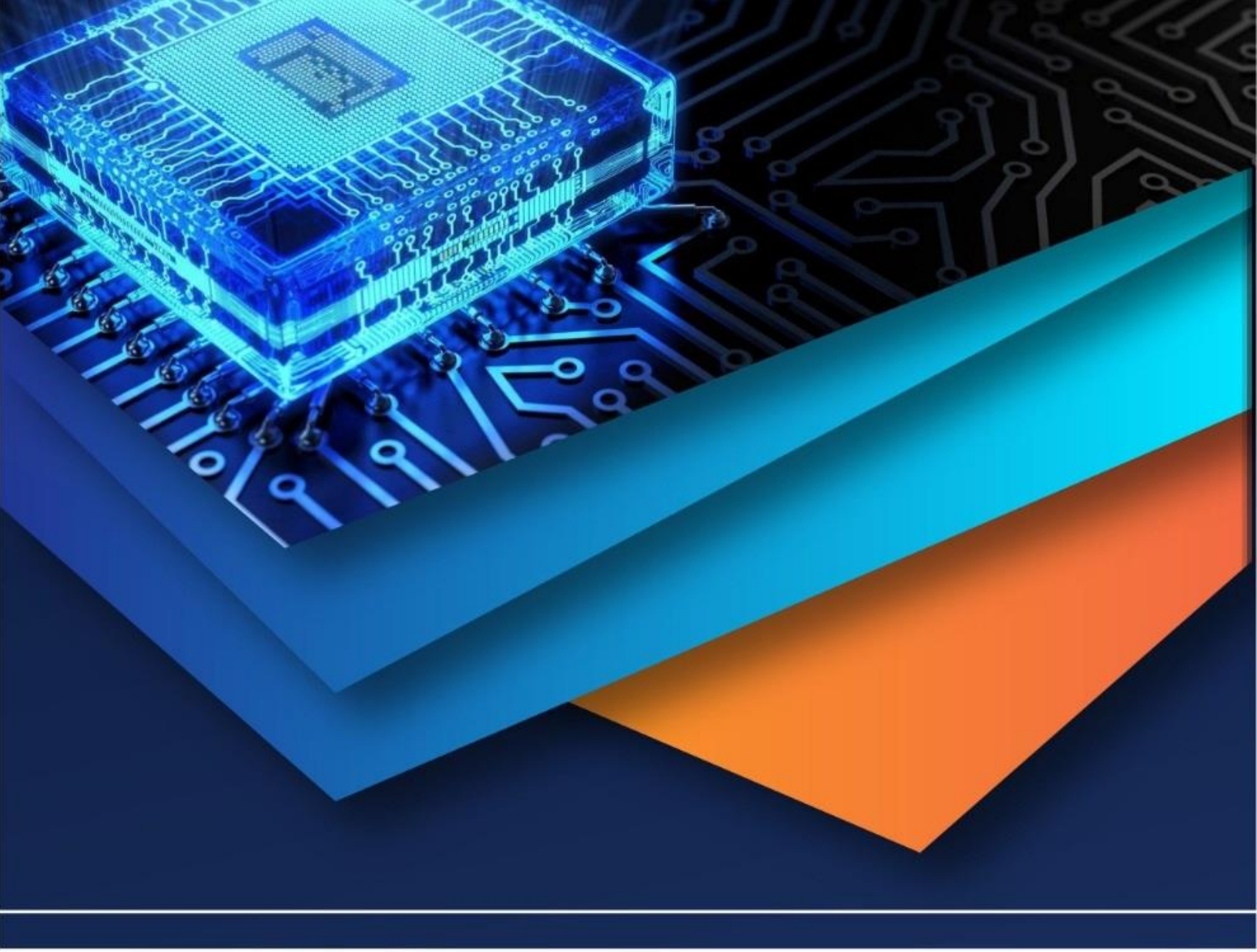
©IJRASET: All Rights are Reserved

The poll also covered the shortcomings of the hate speech detection techniques in use today, with a focus on the difficulties in identifying implicit hate speech, sarcasm, and irony. In order to better capture language variances and cultural subtleties, it also highlighted the need for more representative and varied datasets. In summary, the intrinsic complexity and cultural sensitivity of the Roman Hindi language pose a considerable obstacle for the identification of hate speech in the language. For this job, BERT and CNN models have shown promise; however, because of their better performance in capturing contextual information, BERT models outperform CNN models. To overcome the shortcomings of the present approaches and create more reliable and culturally aware detection models, further research is still required.

REFERENCES

[1] A. Rafae, A. Qayyum, M. Moeenuddin, A. Karim, H. Sajjad,and F. Kamiran, ‘‘An unsupervised method for discovering lexical variations in Roman Urdu informal text,’’ in Proc. Conf.Empirical Methods Natural Lang. Process., vol. 2015, pp. 823– 828.
[2] Z. Sharf and S. U. Rahman, ‘‘Lexical normalization of Roman Urdu text,’’ Int. J. Comput. Sci. Netw. Secur., vol. 17, no. 12, pp. 213–221, 2017.
[3] B. Han and T. Baldwin, ‘‘Lexical normalisation of short text messages: Makn sens a# Twitter,’’ in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Hum. Lang. Technol., 2011, pp. 368–378.
[4] B. Han, P. Cook, and T. Baldwin, ‘‘Lexical normalization forsocial media text,’’ ACM Trans. Intell. Syst. Technol., vol. 4, no.1, pp. 1–27, 2013.
[5] D. Supranovich and V. Patsepnia, ‘‘IHS_RD: Lexical normalization for English tweets,’’ in Proc. Workshop Noisy User-Generated Text, 2015, pp. 78–81.
[6] N. Kaji and M. Kitsuregawa, ‘‘Accurate word segmentation and POS tagging for Japanese microblogs: Corpus annotation and joint modeling with lexical normalization,’’ in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2014.
[7] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, ‘‘A lexicon- based approach for hate speech detection,’’ Int. J. Multimedia Ubiquitous Eng., vol. 10, no. 4, pp. 215–230, Apr. 2015.
[8] C. Hutto and E. Gilbert, ‘‘VADER: A parsimonious rule- based model for sentiment analysis of social media text,’’ in Proc. Int. AAAI Conf. Weblogs Social Media, 2014, vol. 8, no. 1, pp. 216–225.
[9] T.Davidson, D. Warmsley, M. Macy, and I. Weber, ‘‘Automated hate speech detection and the problem of offensive language,’’ in Proc. Int. AAAI Conf. Web Social Media, 2017, vol. 11, no. 1, pp. 512–515.
[10] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan, ‘‘All you need is ‘love’ evading hate speech detection,’’ in Proc. 11th ACM Workshop Artif. Intell. Secur., 2018, pp. 2–12.
[11] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, ‘‘Hate speech detection: Challenges and solutions,’’ PLoS ONE, vol. 14, no. 8, Aug. 2019, Art. no. e0221152.
[12] S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. Khan, and G. Mujtaba, ‘‘Automatic hate speech identification: A comparative study,’’ Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 8, pp. 1–8, 2020.
[13] M. M. Khan, K. Shahzad, and M. K. Malik, ‘‘Hate speech detection in Roman Urdu,’’ ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 20, no. 1, pp. 1–19, Apr. 2021.
[14] M. Di Capua, E. Di Nardo, and A. Petrosino, ‘‘Unsupervised cyber bullying detection in social networks,’’ in Proc. 23rd Int. Conf. Pattern Recognit. (ICPR), Dec. 2016, pp. 432–437.
[15] A. Rodriguez, C. Argueta, and Y.-L. Chen, ‘‘Automatic detection of hate speech on Facebook using sentiment and emotion analysis,’’ in Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIC), Feb. 2019, pp. 169–174.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)