International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 10 | Oct - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Hate Speech Classification Using Machine Learning

{BE Students} Bhagat Ajay Suresh, Jagtap Radhika Suryakant, Bhalerao Snehal Jagdish, More Siddhi Vijay
{Faculty} Dr. Dinesh Bhagwan Hanchate, Prof Swati Mahadev Atole
Department of Computer Engineering, Dattakala Group Of Institution Faculty Of Engineering, Bhigwan-413130, University of Pune, Maharashtra, INDIA.
moresiddhi186@gmail.com

Abstract

Hate speech is a crime that has increased recently, both online and in face-to-face interactions. There are several reasons for this. On the one hand, people are more prone to act hostilely because of the anonymity provided by the internet, and social networks in particular. On the other hand, people's urge to express their opinions online has grown, which has contributed to the proliferation of hate speech. Given how damaging this kind of discriminatory speech can be to society, governments and social media companies can both profit from detection and prevention strategies. Through this survey, we provide a thorough evaluation of the research conducted in the field, which helps to resolve this conundrum.The use of many complex and non-linear models helped with this difficulty, and CAT Boost outperformed the others because it applied latent semantic analysis (LSA) to reduce dimensionality.

Introduction

The prevalence of hate speech has increased in recent years, both in person and online. Numerous things are involved in this situation. The anonymity of the internet makes people more prone to act hostilely, but it also makes people more inclined to express their opinions online, which contributes to the spread of hate speech. Governments and social media companies can benefit from detection and prevention techniques since this kind of discriminatory speech can have a terrible impact on society. We hope that our survey may shed some light on the numerous studies that have been carried out in this field.

Hate speech is defined as any discourse that has the capacity to do harm to an individual or group and that may result in violence, insensitivity, or illogical or inhuman behavior. The prevalence of hate speech on online social media sites like Facebook and Twitter has increased along with their popularity. There is evidence that hate speech is contributing to an increase in hate crimes. As the issue of hate speech gains traction, numerous government-led initiatives are being put into place, such as the Council of Europe's No Hate Speech campaign. The EU Hate Speech Code of Conduct, which all social media platforms are required to sign and follow within 24 hours, is another way it has been put into effect.

Numerous issues that have been brought to light have generated serious concerns about dataset quality, which is what this study attempts to solve. This work also tackles the second problem, which is that before creating an appropriate classifier, the best features for identifying hate speech must be researched and identified.

The most prevalent categories, according to FBI hate crime data, are race, ethnicity, and religion. Datasets typically fit into one of these categories as a result

Terminologies of Hate Speech Classification for Text

1. **Hate Speech:** The expression of hate, prejudice, or discrimination against people or groups based on factors such as race, religion, gender, nationality, sexual orientation, etc.

2.Offensive Language: It is language that tends to be abusive, rude, or profane in nature but does not necessarily cross the boundaries into hate speech.

3.Toxicity: This is a measure of negative or harmful language that may include hate speech, offensive language, bullying, harassment, or insulted words.

4.Classification: In general, classification refers to the process of predicting categories or labels for data. For text, it can be described that the text is being classified into categories such as "hate speech," "offensive," or "neutral."

5.Machine Learning: A form of AI in which systems can learn from data and make predictions. Simple ML techniques applied in text classification include logistic regression, support vector machines, and decision trees.

6.**Natural Language Processing:** A field of AI focused on the interaction between computers and human language. This field enables the analysis and interpretation of human language in several tasks, such as classification.

7.Text Preprocessing: Techniques that are applied to the text data before training in a model like tokenization, stemming or lemmatization, and removal of stop words mainly with the purpose of standardization and cleaning the text.

8.Tokenization: The process through which the text is broken down into words, phrases, or symbols called tokens for further processing

9. **Sentiment Analysis**: A technique used in NLP for analyzing the sentiment from text, usually classified as positive or negative or neutral. It is remarkably different from hate speech but can sometimes be used as an auxiliary tool

10. Precision and Recall: Metrics for Classifying Performance. Precision is the proportion of identified instances that are actually hate speech. Recall is the proportion of correctly identified actual hate speech by the model.

11. F1 Score: F1 score is controlled precision and recall in a balance to evaluate performance with not too much overprediction and underprediction.

Literature survey

A literature survey on hate speech classification for text would summarize the existing research, methodologies, datasets, challenges, and innovations in this domain**.**

1. Introduction to Hate Speech Classification

Overview of what constitutes hate speech and how it differs from offensive language, harassment, and abusive language.The societal need for automated hate speech detection, especially with the rise of social media and online communication. Discuss challenges like defining hate speech consistently across languages, Ethical challenges in creating unbiased, fair algorithms, especially in handling protected categories such as race, religion, and gender.

2.Datasets for Hate Speech Classification

An overview of popular datasets like Hatebase, Twitter datasets, and others used in hate speech detection research. Each dataset is annotated for hate speech, offensive language, or other relevant categories.Many datasets are domain-specific (e.g., Twitter or Facebook), or have inherent biases.

Datasets may also lack diversity in terms of language and geographic regions

3 Preprocessing Techniques

Text Cleaning: Techniques such as lowercasing, removing special characters, and removing stop words.Tokenization, Stemming, and Lemmatization: Breaking text into tokens and standardizing them for consistency in model input.Handling Imbalanced Common methods like oversampling, undersampling, and data augmentation to handle imbalances in hate speech vs. non-hate speech classes

4. Feature Extraction Techniques

Bag of Words (BoW) and TF-IDF: Traditional techniques for extracting word-level features from text.Word Embeddings: Usage of distributed word representations like Word2Vec, GloVe, and fastText, which capture semantic relationships between words.Advanced Embeddings and Contextual Adoption of transformer-based models like BERT, RoBERTa, and DistilBERT, which have shown improved performance by capturing context more accurately.

5.Machine Learning Models for Hate Speech Classification

Classical Approaches: Early approaches involved Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM). These were used with features extracted from BoW, TF-IDF, or simple word embeddings.CNNs and RNNs (LSTMs and GRUs) have been effective in capturing semantic and contextual information in hate speech classification tasks.Models like BERT, GPT, and their derivatives, which leverage pre-trained knowledge, have significantly improved classification performance.Combining machine learning and rule-based methods, or multiple model types, to improve classification by capturing complex nuances in language.

Implementation details

1. Data Collection and Annotation

Collect text data from social media platforms (like Twitter, Facebook), forums, or dedicated hate speech datasets. Kaggle Datasets: Various datasets focused on hate speech and offensive language in multiple languages. If creating a custom dataset, annotate text samples into categories like "hate speech," "offensive language," and "neutral." Annotation should ideally be conducted by multiple annotators to ensure consistency and reduce bias.Clean data to remove duplicates, bot-generated texts, and other noisy entries that may affect model accuracy

.

2. Data Preprocessing

- **Text Cleaning:** Remove special characters, emojis, and URLs. Lowercase all text for consistency, and remove stop words if they don't add meaning for classification.
- **Tokenization:** Split text into individual tokens (words or subwords). Tools like SpaCy or NLTK can help tokenize text.
- **Stemming and Lemmatization:** Reduce words to their base form (e.g., "running" → "run") to avoid redundancy.
- **Handling Imbalance**: Since hate speech is often less prevalent in datasets, apply techniques like:

➢ **Oversampling**: Increase hate speech samples by duplicating or generating synthetic samples.
➢ **Undersampling**: Reduce the number of non-hate speech samples.
➢ **Data Augmentation**: Use techniques like synonym replacement, back translation, or paraphrasing to generate more hate speech samples.

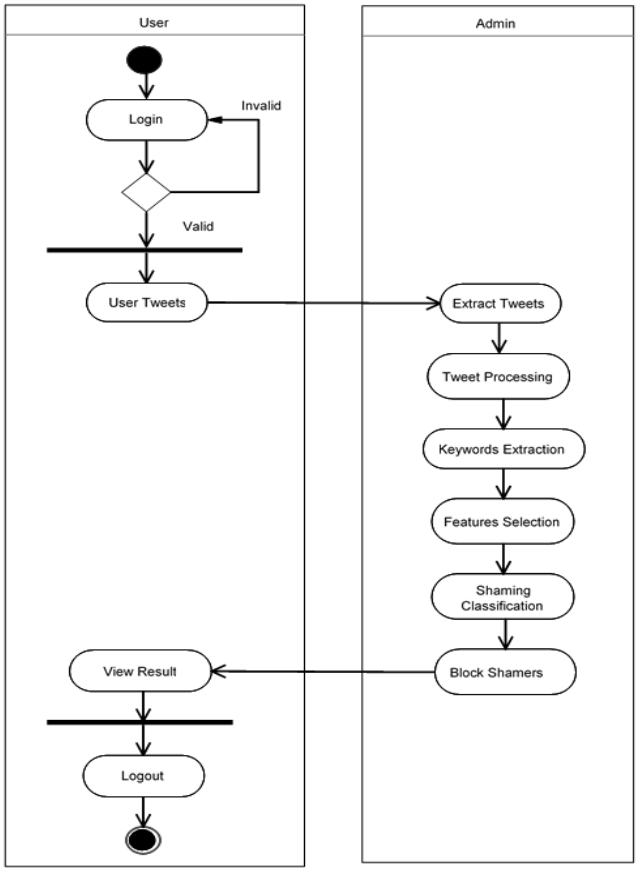

Fig. Activity Diagram

3. Feature Extraction

- **Bag of Words (BoW):** Represent text as word occurrence counts in a matrix.
- **TF-IDF (Term Frequency-Inverse Document Frequency):** A weighted version of BoW, giving higher importance to infrequent words across documents.
- **Word Embeddings:**Word2Vec, GloVe, FastText: Pre-trained embeddings that capture semantic meaning
- **Transformer-based Embeddings:** Using pre-trained models like BERT, RoBERTa, or DistilBERT, which create embeddings based on the context in which words appear.

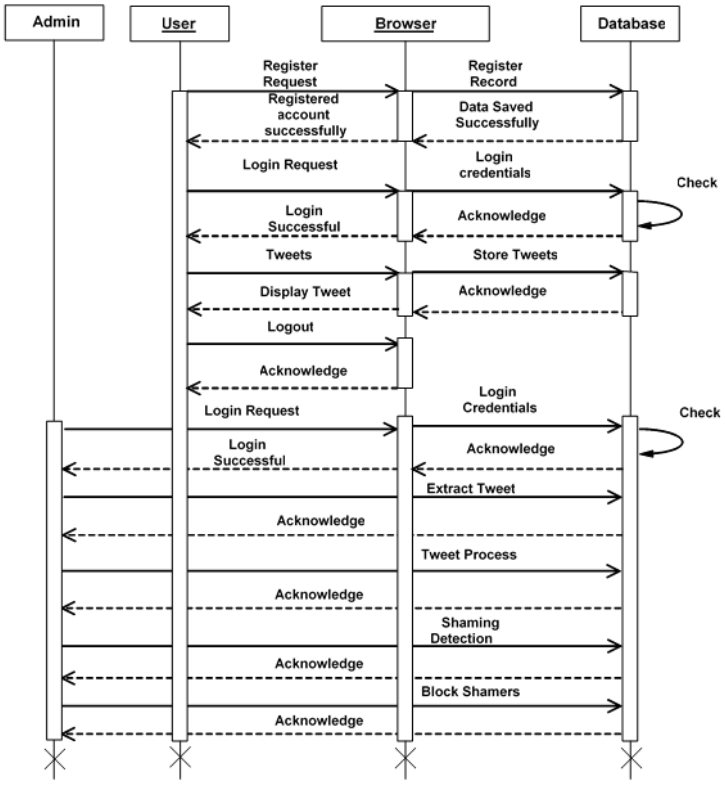

Fig. Sequence Diagram

4. Evaluation Metrics

- **Precision**: Measures how many of the hate speech predictions are correct.
- **Recall**: Measures how well the model identifies all actual hate speech cases.
- **F1 Score**: Balances precision and recall, especially useful in imbalanced datasets.
- **Confusion Matrix**: Provides insights into true positives, false positives, true negatives, and false negatives, helping identify where the model misclassifies.

5. Handling Bias and Fairness

- **Bias Detection**: Test the model on diverse demographic groups to ensure fair performance across races, genders, and other protected groups.

- **Fairness Metrics**: Calculate fairness metrics, such as equalized odds or disparate impact, to assess the model's bias levels.

Results and discussion

Hate speech detection in Tex was ignored in previous technology because there was no survey on automatic detection.In the White Supremacy Forum, there are far more sentences that do not convey hate speech than there are 'hateful' sentences.

It's possible that the increase in the F1-score on the two datasets was influenced by the individual feature (number of) 'Followers', which also improved the subset improvement.These unigrams and patterns can be used as already-built dictionaries not included in the proposed hate speech detection dictionaries for future research projects

Conclusion

After identifying the primary challenges, the multi-class automated hate speech categorization for text problem is solved with significantly better results. It is possible to categorise hate speech into one of ten distinct binary datasets.Each dataset was annotated by a team of experts who followed a set of specific guidelines to the letter. All of the data was evenly distributed across the different datasets. They were also given a boost in terms of subtlety in language. To fill the gap in the field, a dataset like this had to be compiled.

Acknowledgements

This paper would not have been written without the support and encouragement of Asst. Prof. Dr. D.B.Hanchate, guide of BE Dissertation work. Author's special thanks go to all the professors of computer engineering department of DGOI FOE Swami chincholi, for their guidance and for giving her an opportunity to work on Hate Speech Classification on Text using Machine Learning.

References

1. *Hate Speech Explained: A Toolkit, vol. 19, London, U.K., 2015.*
2. *K. Saha, E. Chandrasekharan, and M. De Choudhury, "Prevalence and psychological effects of hateful speech in online college communities," in Proc. 10th ACM Conf. Web Sci., Jun. 2019, pp. 255–264.*
3. *M. Bilewicz and W. Soral, "Hate speech Epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization,"*
   Political Psychol., vol. 41, no. S1, pp. , Aug. 2020.
4. *E. Blout and P. Burkart, "White supremacist terrorism in charlottesville: Reconstructing unite the Right," Stud. Conflict Terrorism, pp. 1–22, Jan. 2021.*
5. *R. McIlroy-Young and A. Anderson, "From 'welcome new gabbers' to the Pittsburgh synagogue shooting: The evolution of gab," in Proc. Int. AAAI Conf. Web Social Media, vol. 13, 2019, pp. 651–654.*
6. *A. Warofka, "An independent assessment of the human rights impact of Facebook in Myanmar," Facebook Newsroom, vol. 5, Nov. 2018.*
7. *T. H. Paing, "Zuckerberg urged to take genuine steps to stop use of Fb to spread hate in Myanmar," Irrawaddy.*
8. *Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in Proc. NAACL StudentRes. Workshop, 2016, pp. 88–93.*
9. *T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," Proc. Int. AAAIConf. Web Social Media, vol. 11, no. 1, pp. 512–515, May 2017*
10. *A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. 5th Int. Workshop Natural Lang. Process. Social Media, 2017, pp. 1–10.*

11. *P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, 2018.*
12. *V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of*
13. *hate speech against immigrants and women in Twitter," in Proc. 13th Int. Workshop Semantic Eval. Vancouver, BC, Canada: Association for Computational Linguistics, 2019, pp. 54–63.*

Author Biblography :

Siddhi Vijay More
Birth Place: Bhiwandi,Thane (2002) (MH- India), received diploma in Information Technology from K.K. Wagh Polytechnic,Nashik. Currently pursuing B.E. in Computer Technology from DGOIFOE.Internship Completed in RPA from ProAzureCompany.Certification in Java, Advance Java and Cyber Security.

Radhika Suryakant Jagtap
Birth Place: Barshi(2004) (MH- India), received dipoma from Government Polytechnic Pune,. Currently pursuing B.E. in computer Technology from DGOI FOE. Cerification in Robotic Process Automation from ProAzure Technology, Kharadi Pune.

Ajay Suresh Bhagat
BirthPlace: Indapur(2001) (MH- India), received diploma in Computer Technology S.B patil polytechnic Indapur. Currently pursuing B.E. in computer Technology from DGOI FOE. Cerification in Robotic Process Automation from ProAzure Technology, Kharadi Pune.

Snehal Jagdish Bhalerao
Birth Place: Bhiwandi (2001)(MH- India), received diploma in Computer Technology from Phivajirao S. Jondhale polytechnic Asangaon. Currently pursuing B.E. in computer Technology from DGOI FOE. Certification in Robotic Process Automation from ProAzure Technology, Kharadi Pune.

Dr.Dinesh Bhagwan Hanchate

Birth- Solapur (MH- India), B.E. Comp. (Walchand College of Engg., Sangli , MH, (India)), M. Tech. Comp. (Dr. Babasaheb Ambedkar Technological University, Lonere, MH (India)). Ph.D. Comp. (SGGSIET, Nanded and SRTMU, Nanded, MH (India)). Former HOD of Comp. and IT. Head PG section, PG teacher. Did STTP, QIP programs sponsored by IIT, Kanpur, AICTE, ISTE, SPPU and UG. Positions: PhD Guide, Dean Industry Institute Interaction Cell.R & D coordinator, Head Media Cell, Editor Art Magazine, Students Development Officer, System Officer, Pune Division Head, Career Katta, GoM DTE, Author, Poet, Currently Professor in DGoI, Daund. Interest in ML, Software Engineering, AI, IR, Math Modelling, Usability Engg., Optimization, Soft Computin g..

Email: dineshbhanchate@gmail.com

Swati Mahadev Atole

Name: Swati Mahadev Atole
Birthplace:Korti(Karmala) (MH- India)
Education: BE computer science and engineering from Karmaveer Bhaurao Patil College of Engineering ,Satara(2013)
Currently pursuing M.E. in Computer Technology from DGOI FOE. Currently working as Assistant Professor (Computer & IT) dept in Dattakala Group of Institutions Faculty of Engineering,Bhigwan.
Email: swati.atole03@gmail.com