Explainable and Debiased Misogyny Identification In Code-Mixed Hinglish using Artificial Intelligence Models

Sargam Yadav[1]

[1]*Dundalk Institute of Technology, Dundalk, A91 K584, Ireland*

Abstract

Hate speech in online platforms has become a significant problem. Women generally face more online harassment, and it is crucial to investigate automated approaches to filter large volumes of data to detect hate speech. In recent years, artificial intelligence models have shown great potential in automatic filtering of hate comments. In this project, a novel dataset for misogyny detection in code-mixed Hinglish will be created from YouTube and Twitter (X) and labelled for coarse-grained and fine-grained misogyny detection. Explainable Artificial Intelligence (XAI) tools will be applied to the resulting models to ensure that they are using logical features for classification. The resulting models will also be evaluated on fairness metrics, and efforts will be made to measure and mitigate unintended biases from all steps of the training process. To date, the literature review for the project has been completed and the data collection and annotation process has been started.

Keywords

Natural language processing, Code-mixed languages, Hinglish, Hate speech detection, Explainable artificial intelligence, Misogyny detection

1. Introduction

In the era of digital communication, social media platforms have become the center of social interaction [1, 2, 3, 4], business, and communication. However, this ease of communication along with the perceived anonymity of the internet has allowed hate, misinformation, and cyberbullying to spread even more rapidly than in the real world. Hate speech is any derogatory speech towards a group of people based on identity factors such as race, gender, nationality, etc. Misogyny is defined as an ingrained prejudice against women. Women are disproportionately targeted on social media platforms, and this harassment has unfortunately led to several real-world consequences [5]. In recent years, machine learning and deep learning models have shown great potential in automatically filtering toxic and hateful content [6] [7]. Studies have delved into using simple machine learning models such as Support Vector Machine (SVM) [8] and Logistic Regression (LR) [7] as well as more complex deep learning models such as Bidirectional Encoder Representations from Transformers (BERT) [9], Recurrent Neural Network (RNN), and Long short-term memory (LSTMs) to automatically detect hate speech in languages such as

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

sargam.yadav@dkit.ie (S. Yadav)

0000-0001-8115-6741 (S. Yadav)

English [10], Italian [7], and more. However, the research effort is scarce for under-resourced and code-mixed languages. Code-mixed languages combine the features of two languages and are a product of multilingual societies. India, with over 467 million active users [11], is home to code-mixed languages such as Hinglish. Hinglish is a portmanteau of Hindi and English, and combines Romanized Hindi with English words.

The decision making process of deep learning models cannot be analysed directly due to its black-box nature. To make models interpretable and trustworthy, XAI techniques can be implemented at the global and local level. Feature attribution methods such as Local Interpretable Model-Agnostic Explanations (LIME) [12] and Shapely Additive explanations (SHAP) [13] can be used to understand which parts of the input affect the final prediction. Artificial intelligence models are trained on real-world datasets, which can unfortunately reflect societal biases [14] [15]. The model can then amplify this bias and further propagate harmful stereotypes. In this project, a novel dataset of Hinglish comments collected from YouTube and Twitter will be annotated for misogyny classification at two levels: binary classification into misogyny and not-misogyny, and multi-class classification of the misogynistic comments into additional labels. Feature vectors will be obtained from the dataset using techniques such as TF-IDF, Word2Vec, etc. Additional features such as length of comments, URLs, etc., will also be considered. Machine learning and deep learning models will be trained on these features and evaluated using standard metrics such as accuracy, F1-score, precision, recall, etc. The best models will be then evaluated using XAI techniques to ensure that the model behaviour aligns with human rationale. Additionally, unintended bias will be measured from the word embeddings and the resulting model, and mitigation strategies will be applied accordingly. A toolkit will be developed from the resulting debiased and explainable model. The rest of the article is structured as follows: Section 2 describes the motivation behind conducting the project as well as the research hypothesis, questions, and objectives. Section 3 covers the relevant literature in the areas of hate speech detection, code-mixed and under-resourced languages, explainability in hate speech classification, and fairness and bias in hate speech classifiers. Section 4 details the methodology used in the project. Section 5 outlines the progress of the project to date. Section 6 discusses the plan for progression of the project, and Section 7 concludes the study.

2. Motivation

Social media platforms have become the epicenter of communication, business, and political discourse. Individuals from different demographic groups can connect with like-minded individuals and share their opinions. However, reports have demonstrated that women politicians and journalists face a significant amount of online gendered trolling and harassment, which can cause them to withdraw from online discussions, effectively silencing their voices [5]. The continued and targeted harassment facilitated by social media can also adversely impact the mental health of victims. Unfortunately, a number of cases of online harassment have resulted in real-life harassment and murders [16]. Therefore, a proactive approach to automatically filtering misogyny in online comments is essential to ensure a safe and inclusive digital space. Previous studies and shared tasks have demonstrated the potential of Natural Language Processing (NLP) techniques in monitoring online content. Therefore, the project will attempt to build on these

findings and provide a novel dataset and a debiased and explainable toolkit to automatically filter misogyny in code-mixed Hinglish comments.

The research objectives, hypothesis, and questions are highlighted in the next sections.

2.1. Research Objectives

The objectives of the project are listed as follows:

1. Review current state-of-the-art in hate speech and misogyny detection, explainable hate speech classifiers, and fairness and bias in NLP to identify knowledge gaps and compare approaches.
2. Implement and compare different statistical and contextual feature embedding techniques.
3. Investigate the explainability of the classifiers through XAI tools and analyse performance on explainability metrics.
4. Analyse model fairness by measuring performance disparity for different demographics.

2.2. Hypothesis and Research Questions

After a comprehensive literature review and preliminary studies, the following research hypothesis has been formulated:

Hypothesis: Artificial intelligence models can be used to detect misogyny in code-mixed Hinglish comments in a fair and explainable manner.

The research questions formulated to explore the hypothesis are as follows:

1. Research Question 1: What are the current benchmark datasets and state-of-the-art approaches for hate speech and misogyny detection in code-mixed Hinglish and other languages?
2. Research Question 2: Which artificial intelligence models and feature vectorization techniques perform the best at misogyny classification on the novel code-mixed Hinglish dataset?
3. Research Question 3: What is the outcome of applying XAI techniques on the misogyny classifier?
4. Research Question 4: What is the outcome of performing bias measurement and mitigation on the misogyny classifier?

3. Related Work

Hate speech is a complex issue that needs to be tackled without infringing on free speech. In recent years, several studies have been conducted to counter hate speech from several perspectives in various languages. The Automatic Misogyny Identification (AMI) @ Evalita 2018 shared task focused on English and Italian datasets of 10000 tweets each and consisted of two subtasks: misogyny identification, and categorization of misogynistic behaviour and identification of the target of such behaviour [7]. The Explainable Detection of Online Sexism task at SemEval-2023 [10] focused on three subtasks: classification into sexist and non-sexist, classification of sexist

posts into 4 categories: 1: threats, plans to harm and incitement, 2: derogation, 3: animosity, and 4: prejudiced discussion, and classification into 11 fine-grained vectors. The best scores were achieved by models such as DeBERTa-v3 and TwHINBERT. Hate speech detection in low-resource languages is a challenging task as social and cultural perspectives are required to understand the nuances of a language. Identification of Conversational Hate-Speech in Code-Mixed Languages ICHCL – 2021 [6] was a subtask at Hate Speech and Offensive Content Identification (HASOC) 2021 shared task that evaluated contextual hate classification in code-mixed Hinglish. The best macro-F1 score obtained was 0.7253 by using a hard voting-based ensemble of 3 transformer-based models: IndicBERT, Multilingual-BERT, and XML-RoBERTa. Bhattacharya et al. [8] presented a dataset of over 25,000 YouTube comments in Indian English, Hindi, and Indian Bangla for aggression and misogyny identification. An SVM classifier using 3-, 5-, and 2- character n-grams yielded the best F1 scores of 0.87, 0.89, and 0.93 for Hindi, Bangla, and English, respectively.

Deep learning models are considered black boxes because their decision-making process cannot be analysed to ensure that they are behaving logically. XAI techniques allow for the exploration and examination of model behaviour to ensure responsible and fair use. DeepHate-Explainer [17] is an explainable approach to hate speech classification in the Indian language 'Bengali'. Sensitivity analysis and layer-wise relevant propagation are used for explainability at a local level, and the faithfulness with respect to comprehensiveness and sufficiency are used at the model level. Artificial intelligence tools are modelled from real-world data which contain biases. Bolukbasi et al. [14] demonstrated that Word2Vec embeddings exhibit gender bias through stereotypes, such as associating 'man' with 'computer programmer' and 'woman' with 'homemaker'. The study introduces two mitigation approaches: hard debiasing and soft debiasing. Nozza et al. [15] evaluate unintended bias from neural network for misogyny classification trained on the dataset in [7] using a synthetic identity test set and AUC-based metrics. Data augmentation has been performed by sampling data from an external corpus to balance the class distribution of identity terms. The literature review suggests that deep learning models perform better at the task than traditional machine learning models due to their complex architectures. Models that score higher on empirical metrics such as accuracy and F1-score, may not be the most explainable or interpretable. Additionally, models can become biased towards identity terms such as 'women', 'wife', etc. Therefore, it is important to explore explainability, and measure and mitigated unintended bias from the model.

4. Research Approach, Methods, Rationale

Figure 1 displays the methodology of the project. A novel dataset for misogyny identification in code-mixed Hinglish will be compiled from YouTube and Twitter (X), and annotated by two annotators at two level: binary classification into Misogyny (MGY) or Not Misogyny (NOT), and a multi-label classification of tweets from the MGY class into labels such as the following: 'Derailing', 'Sexual Harassment and Threats of Violence', 'Stereotyping and Objectification', 'Minimization and Trivialization', 'Whataboutism', 'Religion-based', 'Moral Policing', 'Shaming', and 'Victim Blaming'. Inter-annotator agreement will be calculated using the Cohen's kappa, and the final labels will be decided after deliberation. Where necessary, cleaning and pre-

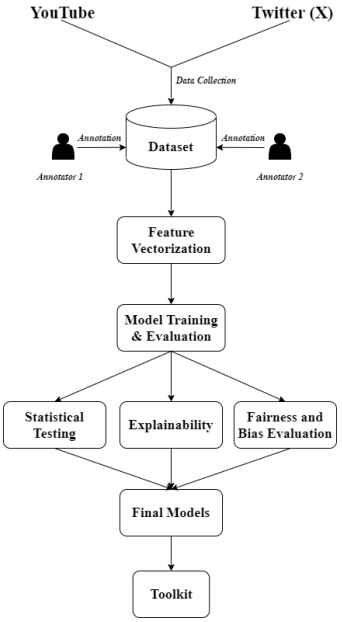

Figure 1: Flowchart of Methodology

processing will be performed by removing punctuation, stopwords, hyperlinks, URLs, etc. Feature vectors will be extracted from the annotated and cleaned dataset using techniques such as Term Frequency - Inverse Document Frequency (TF-IDF) vectorizer, count vectorizer, hashing vectorizer, Word2Vec, GloVe, fastText, etc. Parametric (LR, Naive Bayes, Linear SVM, Perceptron, and more) and non-parametric (K-nearest neighbours, Decision Trees, Random Forests, etc.) machine learning models will be trained and compared. Deep learning models such as RNN, LSTM, Convolutional Neural Networks (CNN), BERT, Generative Pre-Trained Transformer (GPT), Robustly Optimized BERT Pre-training Approach (RoBERTa), etc., will also be implemented. The models will be evaluated on standard evaluation criteria such as F1 score, accuracy, precision, recall, and more. After the training and evaluation step, the best models will be examined using XAI techniques such as LIME and SHAP. To compare all the XAI techniques, a custom evaluation framework will be developed from empirical and human-readable metrics. To measure bias in the feature vectorization step, metrics such as the Word Embedding Association Test (WEAT) will be used to calculate biases in the word embeddings. After the training is completed, the best models will be evaluated for gender bias on synthetic identity test sets. Bias mitigation strategies such as hard debiasing, soft debiasing, data augmentation, etc., will be utilised, and the performance of the debiased models will be compared. After obtaining the final debiased explainable models, a toolkit will be developed to detect misogyny from comments.

5. Preliminary Results and Contributions to date

Since the commencement of the project, considerable progress has been made. The specifics are as follows:

1. Literature Review: The preliminary literature review for the project has been completed. The paper 'Comprehensive Analysis of the Artificial Intelligence Approaches for Detecting Misogynistic Mixed-Code Online Content in South Asian Countries: A Review' [18] was published in the Cyberfeminism and Gender Violence in Social Media journal. An additional survey paper titled 'Exploring Hate Speech Classification in Low-Resource Languages: A Comprehensive Review' is currently under review.
2. Dataset Collection and Annotation: Ethical permission for the data collection and annotation process has been obtained from the host institute. The data collection and annotation process has been started. Approximately 2600 YouTube comments have been collected and annotated by me and my primary supervisor for misogyny identification and categorization of misogynistic behaviour. The paper titled 'Exploratory Data Analysis on Code-mixed Misogynistic Comments.' [19] performs exploratory analysis on the current dataset using techniques such as word cloud visualisations, sentiment polarity, PCA, etc. The Hinglish comments were clustered together. Comments belonging to the class 'MGY' are longer in length, as concluded by previous studies [15].
3. Model Training and Evaluation: The paper 'Hate Speech is not Free Speech: Explainable Machine Learning for Hate Speech Detection in Code-Mixed Languages.' [20] was published and presented at IEEE-ISTAS. The paper compared the performance of machine learning models and 3 feature vectorization techniques: TF-IDF, count vectorizer, and Word2Vec, on the combination of datasets from HASOC-ICHCL 2020 and 2021. LIME was also applied to the best performing model to interpret its results. The paper 'Leveraging Weakly Annotated Data for Hate Speech Detection in Code-Mixed Hinglish: A Feasibility-Driven Transfer Learning Approach with Large Language Models.' explores zero-shot, one-shot, and few-shot learning techniques on a subset of the novel dataset (100 comments).
4. Explainability: A brief exploration of LIME, SHAP, and attention scores was performed in the 'Understanding Interpretability: Explainable AI Approaches for Hate Speech Classifiers.' paper which was accepted and presented at the XAI World Conference 2023. Models were trained on the combined HASOC-ICHCL dataset, and the XAI techniques were applied and analysed.
5. Fairness and Bias: A survey paper titled 'Unveiling the Layers: A Comprehensive Review of Fairness and Bias in Natural Language Processing' has been submitted to a relevant journal and is currently under review. The paper reviews the various sources of biases in NLP systems, fairness criteria and metrics, and bias measurement and mitigation strategies.

6. Next Steps

The project has been divided into the following 4 phases:

1. Data Gathering and Foundational Research (April - July 2024): It consists of the following steps: data scraping (tweets and YouTube comments), data cleaning, and manual annotation. Literature review will be performed regularly to keep track of the current state-of-the-art.
2. Model Implementation and Evaluation (August - December 2024): It consists of the following tasks: pre-processing and feature extraction, model training, and model evaluation and comparative analysis. A large dataset consisting of comments from YouTube and Twitter (X) will be compiled and annotated by the end of July 2024. A DOI will be obtained for the dataset, and it will be made publicly available in accordance with the host institute's open access policy. 1 conference paper will be targeted for EACL 2025.
3. Model Explainability, Fairness, and Bias (January – October 2025): It consists of the following tasks: investigate XAI techniques, evaluate model fairness, evaluate and mitigate unintended biases, analyse results. 1 conference paper will be targeted at XAI-World Conference 2025.
4. Toolkit Development, Writeup, and Completion (November 2025 – March 2026): It consists of the following steps: finalize toolkit, PhD writeup, and PhD examination. 1 conference paper will be targeted for SIGIR 2026.

7. Conclusion and Future Work

Hate speech detection is a highly challenging and nuanced task due to its contextual nature. Previous approaches have achieved a reasonable level of success in differentiating between hate and non-hate content, but still struggle in differentiating between types of hateful behaviour. The project will provide a novel dataset for misogyny detection in code-mixed languages, and evaluate the performance of various machine learning and deep learning models. The models will also be evaluated using XAI techniques such as LIME and SHAP. Additionally, unintended bias will be measured from various stages of the training process and debiasing methods will be implemented. The preliminary literature review for the project has been completed. So far, approximately 2600 comments have been labeled and EDA has been performed on them. Moving forward, the data collection and annotation process will be completed, and model training and evaluation will be performed.

Acknowledgments

This work has been supported by the Regulated Software Research Center at Dundalk Institute of Technology.

References

[1] G. Kaur, A. Kaushik, S. Sharma, Cooking is creating emotion: A study on hinglish sentiments of youtube cookery channels using semi-supervised approach, Big Data and Cognitive Computing 3 (2019) 37.

[2] S. Venkatakrishnan, A. Kaushik, J. K. Verma, Sentiment analysis on google play store data using deep learning, Applications of Machine Learning (2020) 15–30.
[3] S. R. Shah, A. Kaushik, S. Sharma, J. Shah, Opinion-mining on marglish and devanagari comments of youtube cookery channels using parametric and non-parametric learning models, Big Data and Cognitive Computing 4 (2020) 3.
[4] S. Yadav, A. Kaushik, M. Sharma, S. Sharma, Disruptive technologies in smart farming: an expanded view with sentiment analysis, AgriEngineering 4 (2022) 424–460.
[5] A. International, Toxic Twitter – A Toxic Place for Women, https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1-1/, 2024. Accessed: 2024-04-18.
[6] S. Satapara, S. Modha, T. Mandl, H. Madhu, P. Majumder, Overview of the hasoc subtrack at fire 2021: Conversational hate speech detection in code-mixed language, Working Notes of FIRE (2021).
[7] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018) 59.
[8] S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, A. K. Ojha, Developing a multilingual annotated corpus of misogyny and aggression, arXiv preprint arXiv:2003.07428 (2020).
[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[10] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 Task 10: Explainable Detection of Online Sexism, arXiv preprint arXiv:2303.04222 (2023).
[11] W. P. Review, YouTube Users by Country 2024., https://worldpopulationreview.com/country-rankings/youtube-users-by-country., 2024. Accessed: 2024-04-18.
[12] M. T. Ribeiro, S. Singh, C. Guestrin, " Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[13] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[14] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016).
[15] D. Nozza, C. Volpetti, E. Fersini, Unintended bias in misogyny detection, in: Ieee/wic/acm international conference on web intelligence, 2019, pp. 149–155.
[16] S. Chan, "Right-Wing Extremist Convicted of Murdering Jo Cox, a U.K. Lawmaker." New York Times., https://www.nytimes.com/2016/11/23/world/europe/thomas-mair-convicted-murder-jo-cox.html., 2016. Accessed: 2023-02-01.
[17] M. R. Karim, S. K. Dey, T. Islam, S. Sarker, M. H. Menon, K. Hossain, M. A. Hossain, S. Decker, Deephateexplainer: Explainable hate speech detection in under-resourced bengali language, in: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2021, pp. 1–10.
[18] S. Yadav, A. Kaushik, S. Sharma, Comprehensive Analysis of the Artificial Intelligence Approaches for Detecting Misogynistic Mixed-Code Online Content in South Asian Countries: A Review, Cyberfeminism and Gender Violence in Social Media (2023) 350–368.
[19] S. Yadav, A. Kaushik, K. McDaid, Exploratory Data Analysis on Code-mixed Misogynistic Comments, arXiv preprint arXiv:2403.09709 (2024).
[20] S. Yadav, A. Kaushik, K. McDaid, Hate Speech is not Free Speech: Explainable Machine Learning for Hate Speech Detection in Code-Mixed Languages, in: 2023 IEEE International Symposium on Technology and Society (ISTAS), IEEE, 2023, pp. 1–8.