Rejected Cookies @ Decoding Faux-Hate: Predicting Fake Narratives and Hateful Content

Joel D Joy
Center of Data for Public Good
Indian Institute of Science
Bengaluru
djoeldjoy@gmail.com

Naman Srivastava
Center of Data for Public Good
Indian Institute of Science
Bengaluru
srinaman2@gmail.com

Abstract

This paper reports the results of our team for the ICON 2024 shared task Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate). The task aims at classifying tweets in a multi-label and multi-class framework. It comprises two subtasks: (A) **Binary Faux-Hate Detection**, which involves predicting whether a tweet is *fake* (1/0) and/or *hate speech* (1/0, and (B) **Target and Severity Prediction**, which categorizes tweets based on their *target* (Individual, Organization, Religion) and *severity* (Low, Medium, High). We evaluated Machine Learning (ML) approaches, including Logistic Regression, Support Vector Machines (SVM), and Random Forest; Deep Learning (DL) methods, such as Artificial Neural Networks (ANN) and Bidirectional Encoder Representations from Transformers (BERT); and innovative quantum hybrid models, like Hybrid Quantum Neural Networks (HQNN), for identifying and classifying tweets across these subtasks. Our experiments trained and compared multiple model architectures to assess their comparative performance and detection capabilities in these diverse modeling strategies.The best-performing models achieved F1 scores of 0.72, 0.76, 0.64, and 0.54 for the respective labels Hate, Fake, Target and Severity. We have open-sourced our implementation code for both tasks on Github[1].

1 Introduction

The digital era has revolutionized communication through social networks, enabling unprecedented connectivity and information sharing. However, this transformation has also brought significant challenges, notably the rise of hate speech and fake news, which threaten the health of online discourse and social cohesion.

Hate speech, which targets individuals or groups based on characteristics like race, religion, or gender, promotes discrimination, causes emotional harm, and can even lead to violence. The anonymity provided by online platforms worsens the problem, spreading harmful messages, marginalizing vulnerable groups, and fostering hostile environments. Similarly, fake news, which involves false or misleading information, erodes trust, deepens divisions, and disrupts democratic systems. Amplified by the viral nature of social networks, misinformation can lead to social unrest and erode public confidence in institutions. Addressing these issues requires advanced detection methods, including natural language processing, machine learning, and collaborative efforts between tech companies, researchers, and civil organizations. Detection mechanisms must also balance free speech with harm mitigation, incorporating cultural and linguistic nuances.

There have been various approaches to utilising machine learning and deep learning to detect hate speech and fake news. These include Machine Learning Models such as Support Vector Machine (SVM), Random Forest and Logistic Regression (Sreelakshmi et al., 2020; Saumya et al., 2022; Biradar et al., 2024b) and Deep Learning Models such as Artificial Neural Networks, Long-Short-Term Memory Networks (LSTM) (Santosh and Aravind, 2019), Bidirectional Encoder Representations from Transformers (BERT) (Biradar et al., 2021; Nayak and Joshi, 2021; Sreelakshmi et al., 2024; Chopra et al., 2023; Biradar et al., 2022, 2023; Kedia and Nandy, 2021). This study aims to test these Machine Learning and Deep Learning models for the Faux-Hate (Biradar et al., 2024b) Shared Task to predict Hate Speech, Fake News (Task-A), Hate Target, and Severity prediction (Task- B). We further explore Quantum-Hybrid Neural Networks for these tasks.

[1] https://github.com/JoelDJ2002/Faux-Hate.git

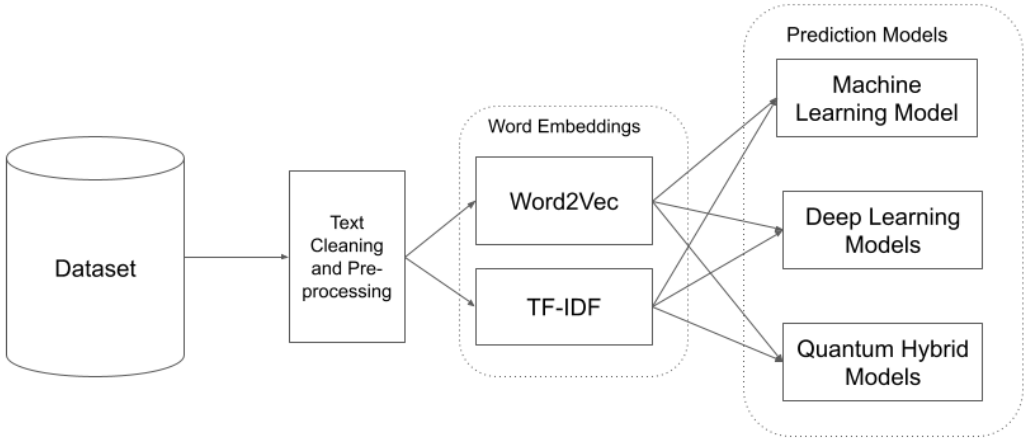

Figure 1: Workflow for the Study

2 Task Description and Dataset Overview

2.1 Task Description

The Faux-Hate shared task (Biradar et al., 2024a,b) aims to address the dual challenge of detecting fake and hate content in social media comments, while also identifying the target and severity of hateful speech. The task is divided into two distinct subtasks:

Task A - Binary Faux-Hate Detection This subtask requires participants to develop a multi-tasking model capable of simultaneously predicting two binary labels for each text sample: (1) *Fake*, which indicates whether the content is fake (1) or real (0), and (2) *Hate*, which determines whether the content constitutes hate speech (1) or not (0). The objective is to enhance the efficiency and accuracy of models by addressing these related tasks jointly.

Task B - Target and Severity Prediction In this sub-task, participants are tasked with building a model that predicts two categorical labels for each text sample: (1) *Target*, which identifies the focus of hateful speech as an Individual (I), Organization (O) or Religion (R), and (2) *Severity*, which categorizes the intensity of the hate speech into Low (L), Medium (M), or High (H). This subtask emphasizes a deeper understanding of the context and impact of hateful content.

3 Methodology

We followed the traditional stages of a Machine Learning experiment illustrated in Fig.1 which includes preprocessing, model training, and evaluation.

3.1 Preprocessing

The preprocessing stage involves preparing the textual data to ensure it is clean and suitable for modeling. Initially, text cleaning is performed to remove special characters, account handles (e.g., @username), and emojis, leaving only the essential text content. The cleaned text is then tokenized into individual words or tokens, facilitating further processing. To represent the text numerically, two embedding techniques are applied: Word2Vec, which generates continuous vector representations based on semantic context, and TF-IDF (Term Frequency-Inverse Document Frequency), which creates weighted vectors that reflect word importance within the corpus. These word embeddings serve as input features for the models.

3.2 Model Training

The study evaluates three categories of models: baseline machine learning models, deep learning models, and quantum hybrid models.

3.2.1 Machine Learning Models

We trained three machine learning models, namely Random Forest, Logistic Regression, and Support Vector Machine. All three models are trained using tf-idf and Word2Vec as embeddings.

3.2.2 Deep Learning Models

Deep learning models are employed to capture complex relationships within the data. An Artificial Neural Network (ANN) is constructed as a feed-forward network with multiple layers. Additionally, BERT (Bidirectional Encoder Representations from Transformers), a transformer-based model, is fine-tuned to the dataset to capture contextual word

relationships between preceding and succeeding words. For training ANNs, Word2Vec and TF-IDF are used as input features, whereas for BERT, we used the pre-trained BERT embeddings (Devlin et al., 2019). For task 1 we use an ensemble of two models, predicting "Hate" and "Fake" separately. Figure 2 depicts the proposed BERT model for task 2, which aims to predict "Target" and "Severity" using a single model.

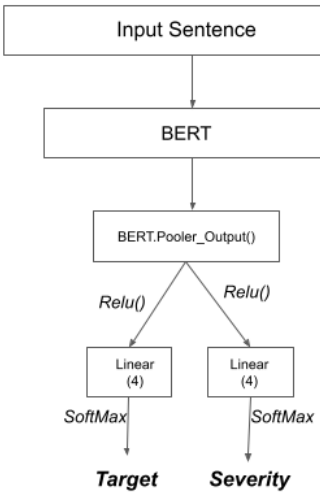

Figure 2: BERT Model for Predicting Target and Severity

3.2.3 Quantum-Hybrid Models

Quantum Hybrid Neural Networks (HQNN), integrate classical neural architectures with quantum components. By leveraging the principles of quantum computing, including superposition and entanglement, these models explore high-dimensional representations to enhance learning capabilities. To create an HQNN we replace the last hidden layer of the ANN with a Quantum Circuit. Figure 3 shows the quantum circuit used for the HQNN. We use a combination of Angle Embedding (Lloyd et al., 2020; Schuld and Killoran, 2019) to encode classical data into a quantum-suitable format, followed by "Basic Entangling Layer" from pennylane (Bergholm et al., 2018; Asadi et al., 2024), which introduces quantum phenomenon such as entanglement into the model(Zaman et al., 2024). We use word2vec as the input feature for the HQNNs.

4 Results and Discussions

In our experiments, we compared various methods for generating word representations, each with different strengths in capturing semantic relationships

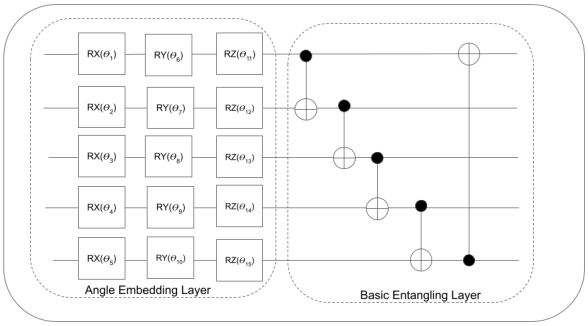

Figure 3: Quantum Circuit for HQNN

within textual data. Word2Vec embeddings outperformed TF-IDF in generating meaningful word representations. BERT (Bidirectional Encoder Representations from Transformers) demonstrated superior performance compared to both Word2Vec based ML models as shown in Table 1. BERT's architecture is designed to capture the bidirectional context of words, allowing it to understand the nuanced meanings of words based on their surrounding text. This contextual awareness enables BERT to outperform previous models by representing words dynamically, depending on their position within a sentence. It has been observed that despite the remarkable theoretical capabilities of Hybrid Quantum Neural Networks (HQNNs), they had scores comparable to ANN, but did not surpass BERT. This may be attributed to the limited number of "qubits" available for training the model, as well as the lack of crucial hyperparameter tuning.

5 Conclusion

This paper describes various approaches used in our approach for the Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (FauxHate) dataset. We investigated the effectiveness of different embedding techniques, demonstrating that word2vec embeddings outperformed TF-IDF in capturing semantic relationships. Furthermore, BERT-based models achieved superior performance, leveraging their contextual understanding to outperform other approaches. It is interesting to note that even though ANN has good learning capabilities, it could not outperform the Machine Learning models. Similarly, Hybrid Quantum Neural Networks (HQNN) showed capabilities similar to ANN, they could not surpass the results obtained by the best-performing BERT-based models. Future work will focus on refining these models and exploring their applicability to similar tasks.

Label	Random Forest		Logistic Regression		SVM		ANN		BERT		HQNN	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
HATE	0.69	0.64	0.67	0.50	0.68	0.51	0.69	0.58	**0.73**	**0.72**	0.69	0.56
FAKE	0.77	**0.77**	0.76	0.76	0.75	0.75	0.75	0.75	**0.79**	0.76	0.52	0.34
TARGET	0.58	0.48	0.55	0.38	0.53	0.32	0.12	0.11	**0.67**	**0.64**	0.09	0.07
SEVERITY	0.49	0.40	0.49	0.37	0.44	0.35	0.29	0.25	**0.58**	**0.54**	0.21	0.16

Table 1: Accuracy and F1 scores of various models with Word2vec.

References

Ali Asadi, Amintor Dusko, Chae-Yeun Park, Vincent Michaud-Rioux, Isidor Schoch, Shuli Shu, Trevor Vincent, and Lee James O'Riordan. 2024. Hybrid quantum programming with pennylane lightning on hpc platforms. *arXiv preprint arXiv:2403.02512*.

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, B Akash-Narayanan, Ali Asadi, et al. 2018. Pennylane: Automatic differentiation of hybrid quantum-classical computations. *arXiv preprint arXiv:1811.04968*.

Shankar Biradar, Sai Kartheek Reddy Kasu, Sunil Saumya, and Md. Shad Akhtar, editors. 2024a. *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In *2021 IEEE international conference on big data (Big Data)*, pages 2470–2475. IEEE.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2022. Fighting hate speech from bilingual hinglish speaker's perspective, a transformer-and translation-based approach. *Social Network Analysis and Mining*, 12(1):87.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2023. Combating the infodemic: Covid-19 induced fake news recognition in social media networks. *Complex & Intelligent Systems*, 9(3):2879–2891.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2024b. Faux hate: unravelling the web of fake narratives in spreading hateful stories: a multi-label and multi-class dataset in cross-lingual hindi-english code-mixed text. *Language Resources and Evaluation*, pages 1–32.

Abhishek Chopra, Deepak Kumar Sharma, Aashna Jha, and Uttam Ghosh. 2023. A framework for online hate speech detection on code-mixed hindi-english text and hindi text in devanagari. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5):1–21.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Kushal Kedia and Abhilash Nandy. 2021. indicnlp@ kgp at dravidianlangtech-eacl2021: Offensive language identification in dravidian languages. *arXiv preprint arXiv:2102.07150*.

Seth Lloyd, Maria Schuld, Aroosa Ijaz, Josh Izaac, and Nathan Killoran. 2020. Quantum embeddings for machine learning. *arXiv preprint arXiv:2001.03622*.

Ravindra Nayak and Raviraj Joshi. 2021. Contextual hate speech detection in code mixed text using transformer based approaches. *arXiv preprint arXiv:2110.09338*.

TYSS Santosh and KVS Aravind. 2019. Hate speech detection in hindi-english code-mixed social media text. In *Proceedings of the ACM India joint international conference on data science and management of data*, pages 310–313.

Sunil Saumya, Vanshita Jha, and Shankar Biradar. 2022. Sentiment and homophobia detection on youtube using ensemble machine learning techniques. In *FIRE (Working Notes)*, pages 92–99.

Maria Schuld and Nathan Killoran. 2019. Quantum machine learning in feature hilbert spaces. *Physical review letters*, 122(4):040504.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

K Sreelakshmi, B Premjith, and KP Soman. 2020. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744.

Kamila Zaman, Tasnim Ahmed, Muhammad Kashif, Muhammad Abdullah Hanif, Alberto Marchisio, and Muhammad Shafique. 2024. Studying the impact of quantum-specific hyperparameters on hybrid quantum-classical neural networks. *arXiv preprint arXiv:2402.10605*.