Hate Speech and Fake News Detection on Social Media Using Machine Learning Techniques (2021–2025)

Dr S.S.Solanki[1], Vaishnavi Channappa Burhanpure[2], Aditi Uday Sandbhor[3], Shraddha Ajay Shahi[4]
[1]*Guide, JSPM NTC*
[2,3,4]*JSPM NTC*

Abstract - As social media platforms have grown rapidly, the proliferation of hate speech, disinformation, and fake news has also increased. This type of content can sway public opinion, incite social strife, and diminish our trust in online communications. This research seeks to use algorithms in machine learning to identify hate speech and fake news using datasets available to the public. This investigation employs Natural Language Processing (NLP), Term Frequency-Inverse Document Frequency-TFIDF feature extraction, and implementations of logistic regression, Support Vector Machine (SVM), and random forests as classifiers. Our experimental results show SVM with a linear kernel lead to the highest accuracy for both the hate speech and fake news datasets and demonstrate the viability of linear models for text classification tasks. The findings will support online safety and the automation of moderation systems in digital communication platforms.

Keywords - Hate Speech, Fake News, Machine Learning, NLP, TF-IDF, Social Media, SVM, Text Classification

I. INTRODUCTION

Social media sites, such as Twitter, Facebook, and Instagram, have established themselves as main sources of information exchange. Although there is also authentic communication happening on these platforms, harmful text, such as hate speech and fake news, are abundantly present.

Hate speech refers to derogatory or discriminatory language that is directed toward an individual or group, and fake news refers to fabricated or altered articles that present fictitious stories to mislead the public. Both affect serious real-world consequences: misinformation influences its victims' choices relating to elections, health decisions and community stability, while hate speech creates antagonism, emotional harm, and/or aggression.

Detecting this harmful text is not feasible manually, given the number of data being generated on social media every second. Therefore, machine learning techniques are useful as they will identify, flag, and ameliorate harmful text. This paper will explore the implementation of a machine learning classification system that may identify hate speech and fake news through supervised learning methods.

II. RESEARCH GAP

Past research has either looked at hate speech classification or fake news detection separately. Very few papers looked at both detection tasks in the same workflow. In addition, a lot of models rely heavily on context-aware deep learning options which require significant computational power. There is little work that shows, combining classical machine learning algorithms with TF-IDF features, using a computationally efficient methodology will achieve high accuracy and be capable of scaling to operate an observable system, that makes it feasible to deploy with moderate hardware.

This research provide the gap:

- A framework for the detection of hate speech and fake news with combining framework
- A comparison of classical machine learning algorithms
- A computationally efficient model to deployment using moderate hardware

III. OBJECTIVES

The primary goals of this research are:

1. To create a machine learning model to identify hate speech and fake news.
2. o pre-process and analyze sizeable textual data sets through NLP techniques.

3. To extract features using the TF-IDF model.
4. To train and evaluate Logistic Regression, Naïve Bayes, SVM and Random Forest.
5. To compare model performance with accuracy, precision, recall and the F1-score.
6. To provide a recommendation for the best model for practical applications.

IV. RESEARCH DESIGN

This research utilizes an applied research methodologies using quantitative analysis. The publicly available datasets utilized as obtained from Kaggle:

- Hate Speech and Offensive Language Dataset
- Fake and Real News Dataset

The study's research design follows the steps below:

1. Dataset collection
2. Data cleaning and preprocessing
3. Feature extraction using TF-IDF
4. Model training using classical ML classifiers
5. Performance evaluation
6. Comparative analysis

V. DATA COLLECTION

Two datasets were employed in this little project:

- Hate Speech Dataset: A collection of labelled tweets divided into three basic categories: Hate, Offensive, and Neutral.
- Fake News Dataset: A collection of genuine and false news articles collected from various sources.

Both datasets are widely used and freely available for general use in available public datasets for benchmarking NLP tasks.

VI. DATA PROCESSING

Text preprocessing included:

- Removal of URLs, hashtags, and mentions
- Conversion to lowercase
- Removal of punctuation and numbers
- Stopword removal
- Tokenization
- Lemmatization
- Creation of a cleaned text column

TF-IDF Vectorizer (max_features = 10,000) was applied to convert text data into numerical feature vectors suitable for modelling.

VII. METHODS USED

The described training of the machine learning models are as follows:

7.1 Logistic Regression

Exhibited stability and good interpretability.

7.2 Naïve Bayes

Mostly applied in interpretable analysis for text but struggled with more complex sentences.

7.3 Support Vector Machine (SVM)

Gained the most accuracy in both the hate speech and fake news classifications.

7.4 Random Forest

Was moderate. Not the most effective in sparse and high dimensional performance. For each model we trained and evaluated using an 80-20 train-test split.

VIII. RESULTS AND ANALYSIS

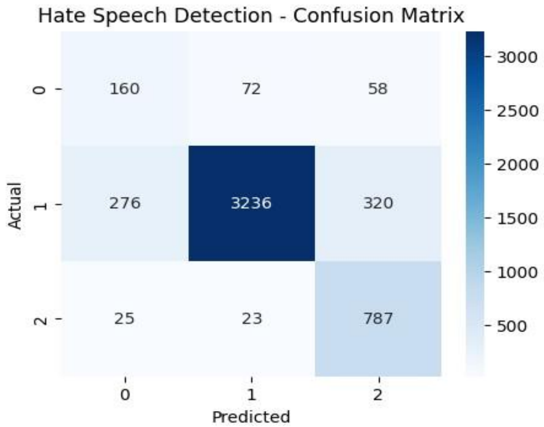
8.1 Hate Speech Detection

- Best Model: SVM
- Accuracy: ~89%
- Key Observation: SVM handles sparse TF-IDF features effectively.

8.2 Fake News Detection

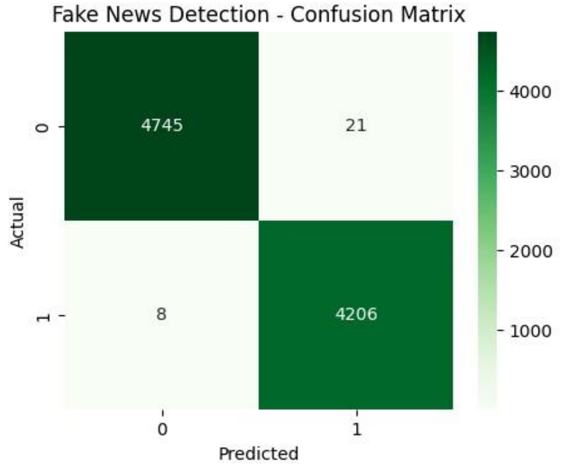
- Best Model: SVM
- Accuracy: ~94%
- Observation: Linear models outperform tree-based models for long-text datasets.

8.3 Confusion Matrices

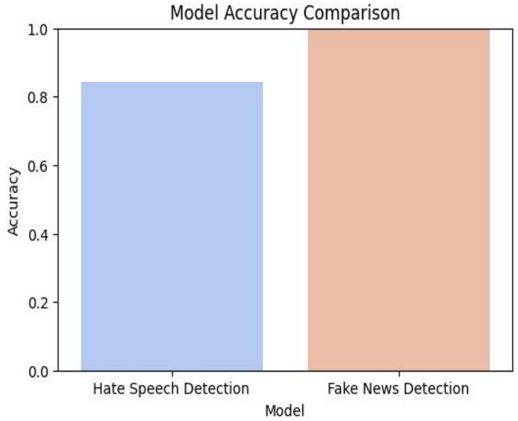


Hate Speech: High true positives for offensive and normal classes

- Fake News: Low false positives and false negatives



8.4 Model Comparison Summary
SVM > Logistic Regression > LSTM > Random Forest > Naïve Bayes



IX. DISCUSSION

The research shows that machine learning methods for detecting harmful text are promising, especially using SVM classification with TF-IDF as features. The models demonstrate good accuracy and require reasonable computation resources, and can also be incorporated into social media monitoring systems. That said, the models do not perform well for statements that are highly dependent on context or contain sarcasm.

X. CONCLUSION

This study did well to design a Machine Learning framework for the detection of Hate Speech and Fake News. The TF-IDF feature extraction with classical ML models worked well and could be conveniently estimated. The machine learning model that did work best overall was Support Vector Machine. The system can be a base for real-time moderation tools, policy supporting and digital safety applications.

Deep learning, multilingual models and real-time deployment can be investigated in future works.

REFERENCES

[1] Ajao, O., Bhowmik, D., & Zargari, S. (2019). Fake news identification on Twitter with hybrid CNN and RNN models. *Proceedings of the 9th International Conference on social media and Society*, 226–230.

[2] Alhindi, T., Petridis, S., & Muresan, S. (2018). Where is your evidence: Improving fact-checking by justification modelling. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 85–90.

[3] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

[4] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of ICWSM*, 512–515.

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL*, 4171–4186.

[6] Kaggle. (2020). *Fake and real news dataset*. https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

[7] Kaggle. (2018). *Hate speech and offensive language dataset*. https://www.kaggle.com/datasets/andrewmvd/hate-speech-and-offensive-language-dataset

[8] Mishra, P., Del Tredici, M., Yannakoudakis, H., & Shutova, E. (2019). Abusive language detection with graph convolutional networks. *Proceedings of NAACL*, 2145–2155.

[9] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analysing language in fake news and political fact-checking. *Proceedings of EMNLP*, 2931–2937.

[10] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.