Hope Speech Identification for Tamil

V. Mathusha and E.Y.A Charles
Department of Computer Science, University of Jaffna, Sri Lanka
mathuvetha1212@gmail.com

ABSTRACT

This research investigates the identification of hope speech in Tamil YouTube comments using Natural Language Processing (NLP) techniques. We compared the performance of various machine learning models, including TF-IDF + SVM, Word Embedding + Linear SVM, and Convolutional Neural Networks (CNN), to classify comments into two categories: hope speech and non-hope speech. The CNN model achieved the highest accuracy of 73%, demonstrating the effectiveness of deep learning approaches in this context.

INTRODUCTION

Hope speech encompasses expressions that convey positivity and encouragement in online discourse. Identifying such speech is essential for fostering supportive digital communities [1]. In Tamil, the complexity of the language presents unique challenges for recognition. This study aims to develop a framework for accurately classifying hope speech using natural language processing techniques, contributing to sentiment analysis and promoting positive communication in digital spaces.

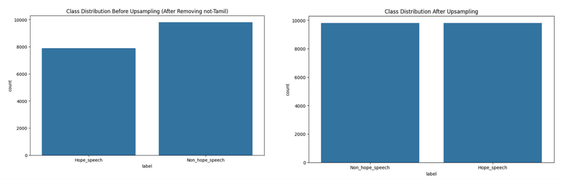
DATASET

- The dataset comprises 20,198 Tamil comments sourced from the Hugging Face website.
- To enhance the quality of the dataset, comments labeled as "not Tamil" were removed, allowing the focus to be solely on the relevant Tamil-language content.

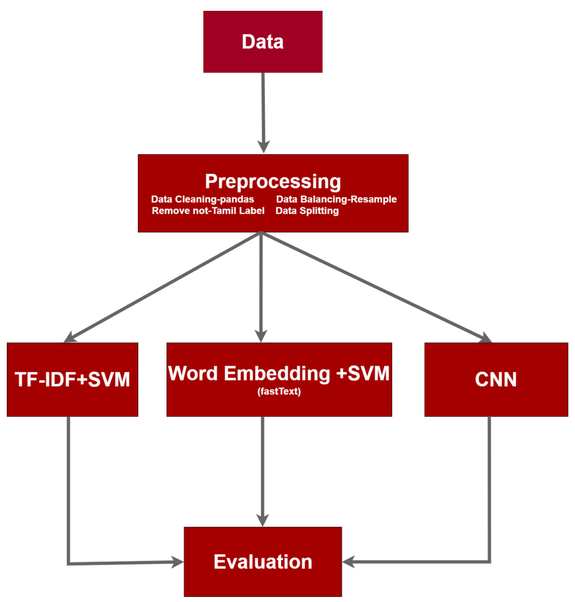
	Text	Label
1	Idha solla ivalo naala	Non_hope_speech
2	இன்று தேசிய பெண் குழந்தைகள் தினம்.. பெண் குழந்தைகளை போற்றுவோம்..அவர்களை பாதுகாப்போம்..	Hope_speech
3	நண்பா நம்ம வீடியோவும் பாருங்க கண்டிப்பா பிடிக்கும் பிடித்தால் support பண்ணுங்க	Hope_speech
4	இந்த மாதிரி பிரச்சினை இல்ல நம்ம வாழனுமா sollu bro	Non_hope_speech

```
Class distribution before upsampling:
label
Non_hope_speech    9816
Hope_speech        7899

Class distribution after upsampling:
label
Non_hope_speech    9816
Hope_speech        9816
```



METHODOLOGY



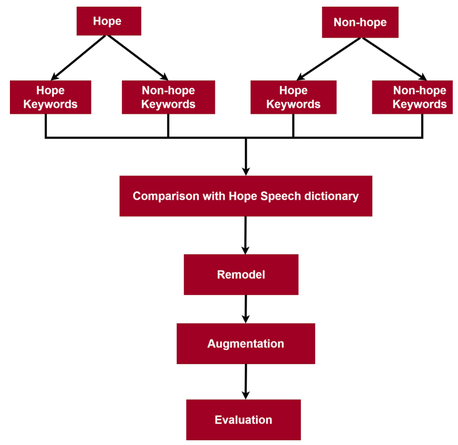
RESULT

Model	Overall Accuracy	F1-score (Hope speech)	F1-score (Non-hope speech)
TF-IDF + SVM	70%	0.70	0.70
Word Embedding + SVM	61%	0.57	0.64
CNN	73%	0.74	0.72

True Label (Actual Data)	Predicted Label	Non-Hope Speech	Hope Speech
Non-Hope Speech (1990)	Non-Hope Speech	1362	628
Hope Speech (1937)	Hope Speech	429	1508

PROPOSED MODEL

- The model analyzes predictions, identifying correct hope speech keywords and misclassified non-hope speech and compared them against a curated Tamil hope speech dictionary for refinement.
- Future improvements include expanding the dictionary, enhancing the model with better architectures, and applying data augmentation for greater accuracy.



CONCLUSION

Our study on hope speech detection in Tamil text demonstrates that Convolutional Neural Networks (CNNs) slightly outperform traditional machine learning methods, achieving 73% accuracy compared to TF-IDF + SVM, Word Embedding + SVM. This suggests that deep learning approaches can effectively capture the nuances of hope speech.

REFERENCES

[1] Chakravarthi B.R. (2020). HopeEDI: A multilingual hope speech detection dataset. CMPOEPESM Workshop, 41-53.
[2] Ziehe, S., et al. (2021). GCDH@ LT-EDI-EACL2021: XLMRoBERTa for hope speech detection. Workshop on LT-EDI, 132-135.
[3] Kumar, A., et al. (2022). Ensemble model for hope speech detection from YouTube comments. 2nd LT-EDI Workshop, 223-228.