**Fake News Detection**

**19ELC381 Open Lab**

***Open Lab Idea Review Report***

*Submitted by*

**CB.EN.U4ELC19004**

**CB.EN.U4ELC19036**

**CB.EN.U4ELC19047**

**CB.EN.U4ELC19052**

**Aditya Narasimhan Sampath**

**Pradhumna Guruprasad**

**Salai Sanjay S**

**Shreyas Nagesh**

April 28, 2022

**2021 - 2022**

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM, COIMBATORE 641112.

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

# Abstract

News tweets and social media posts that report what is happening have become an important real-time information source. In India these are majorly reported in Indian regional languages. Due to COVID these sources of news have grown exponentially. Incidents that are first reported by the individuals on the ground than the media houses. This is where the issue arises. The factual integrity of such news is unknown. The project aims on a comparative study between the various preprocessing models and choose the best among the models present for Indian languages like Hindi. We intend to build our method for tokenization, stop-word removal and lemmatization. This is to optimize it for languages other than English and extract the best performance possible.

# Methodology

Fake news is very widespread nowadays with the advent of social media everywhere. This is a serious problem and has to be addressed in all languages possible. Therefore, this project aims to develop efficient methods to process texts of Indian regional languages. The same methods which apply to English need not work great with other languages. Thus, a comparative study to see which methods best suit what language is performed in this project. According to the language, it is proposed to suggest and improve a method that works best for that language.

First a fake news labelled dataset in Hindi is collected. Pre-processing is performed on this dataset to train the models.

A total of 4 pre-processing methods and 3 machine learning models are explored to see which produces the best result. The pre-processing methods include – TF-IDF, fastText, iNLTK and IndicBERT. The machine learning models explored are – SVM, RCNN and Transformers.

TF-IDF is one of the basic methods which is used to quantify the importance or relevance of words in a document amongst a collection of documents. It is based on only the corpora. fastText, iNLTK and IndicBERT are libraries for word embeddings to help represent words efficiently in the form of numbers. They are trained on numerous languages, so they can also interpret grammar of languages other than English.

3 machine learning models are chosen to see which can perform best. SVM is a classical machine learning model to provide a baseline. RCNN and transformers are the latest advancements in the field of NLP which perform text classification. These models would help classify the text as fake or not.

# Materials Required

|  |  |  |
| --- | --- | --- |
| Software | Price | Availability of Lab |
| Keras | - | Embedded Systems Lab |
| fastText | - | Embedded Systems Lab |
| iNLTK | - | Embedded Systems Lab |
| Tensorflow | - | Embedded Systems Lab |
| IndicBERT | - | Embedded Systems Lab |
| BERT | - | Embedded Systems Lab |

# References

1. <https://github.com/mohit19014/Hindi-Hostility-Detection-CONSTRAINT-2021>
2. <https://ieeexplore.ieee.org/document/9641517>
3. <https://ieeexplore.ieee.org/document/8665593>
4. <https://ieeexplore.ieee.org/document/8777343>
5. <https://ieeexplore.ieee.org/document/9580073>