Fake News Detection

Aditya Sampath – CB.EN.U4ELC19004

Manit Maini – CB.EN.U4ELC19029

Pradhumna Guruprasad – CB.EN.U4ELC19036

Shreyas Nagesh – CB.EN.U4ELC19052

**Amrita Vishwa Vidyapeetham, Coimbatore**

**Abstract**

News tweets that report what is happening have become an important real-time information source. Incidents happening in different corners of the globe are first reported by individuals on the ground than the media houses. This is where the issue arises. We cannot be sure how many of these tweets are true and how many are fake. This project proposes to use Semantic Role Labelling (SRL) and machine learning models to help classify and categorize news articles based on whether they are true or not. A comparative study is done between different machine learning and deep learning models to see which works best for the case of predicting fake news.

**Fake News Detection**

Fake news is becoming more prevalent in today's political scene. This issue needs to be fixed more than ever due to the increasing reliance of people of digital media. News is thrown around so lightly that it is hard to distinguish between real and fake. This experiment aims to establish an end to end working application to detect whether the news is real or fake. To train the model, a dataset with labels denoting whether a piece of news is true, or fake is collected. The data is then cleaned and given to the model.

**Methodology**

A labelled dataset of tweets in English consisting of true and fake news was collected. As we can’t directly use this data in our ML model, we first preprocess the text so that it makes more sense to the algorithm. To make sense of a sentence, it must first be separated into its constituent words. So, tokenization is done first. After this we remove the stop words which are words that don’t convey any meaning or add anything additional to the meaning of a sentence. Since a single word can appear in many forms but ultimately conveys the same meaning, we need to strip the word down to its most basic form that conveys the same meaning. This process is lemmatization. It helps to normalize the different forms of words. Once the data has been preprocessed it is put back into its sentence form but this time with the lemmatized words. This data is split into train and test for the model to learn. The split ratio is 80:20. 80% for training and 20% for testing. After experimenting with different models like logistic regression and standardizing and normalizing data. The CNN starts with a sequential layer followed by a embedding layer. Embedding layer takes care of representing words as word embeddings. This is followed by a pair of convolutional and Average Pooling Layers. To each layer a dropout percentage is added so that overfitting of the model does not occur. The entire CNN outputs a probability value between 0 and 1. A value less than 0.5 denotes that the news is fake while a value greater than 0.5 means that it is true. The block diagram of this entire model can be seen in *Figure 1*.

## Results

4 different variations of models were experimented with. Of the 4, the best performing model was the CNN without any standardization or normalization. The best accuracy was obtained at 76% and the area under ROC curve was 0.829. The results are expressed in a tabulated form in *Table 1*. The ROC curve is plotted in *Figure 2*.

### Conclusion.

Over the last few years, the issue of fake news and its effects on society has attracted increased attention. In the fake news detection issue, the problem of predicting and classifying data needs to be validated using training data. Therefore, a method of fake news detection via the implementation of CNN is proposed in this project. We have used machine learning models to help classify and categorize news articles based on whether they are true or not. We used 4 different approaches to solve the problem and finally concluded that CNN is the best fit for our model, we applied CNN to the tokenized sentences and ran it for 25 epochs. Once this was done, we got the best epoch set which delivered the highest accuracy to predict the fake news. After implementing the proposed method, we evaluated the performance of the proposed method on different test cases. The simulation results showed that the proposed method achieved the required outcomes for detecting fake news.

**References**

# [Shokravi, 2022, Practical Explanation of NLP for Fake News Detection, Medium.com](https://medium.com/geekculture/practical-explanation-of-natural-language-processing-for-fake-news-misinformation-detection-525e26b644c1)

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Tables

Table 1

Results tabulated

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression | 0.70123 |
| CNN | 0.7615 |
| CNN + normalization | 0.5821 |
| CNN + standardization | 0.5801 |

**Figures**

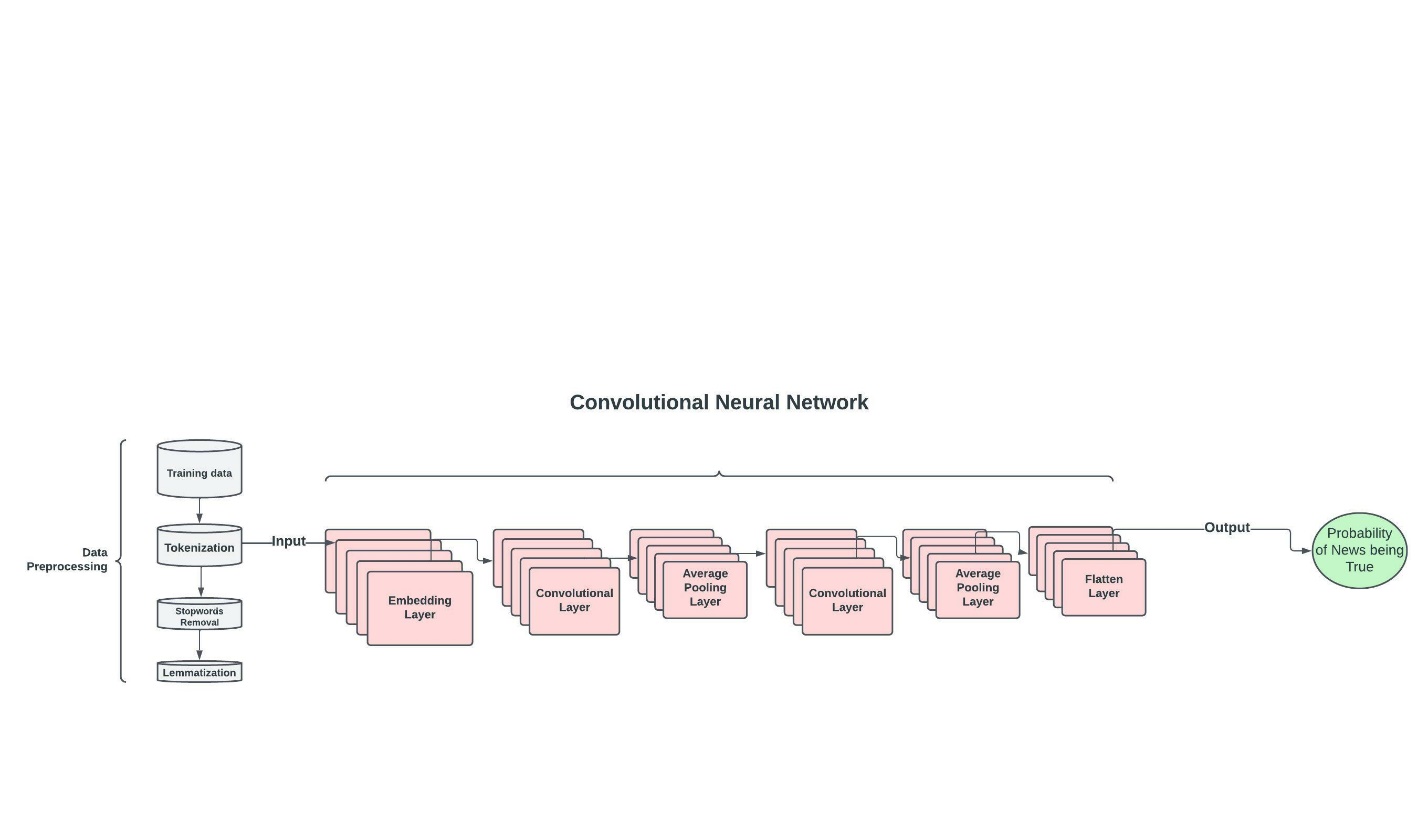


Figure 1: Block diagram of methodology

Chart, line chart

Description automatically generated

Figure 2: ROC curve