AI Course

Team Project Final Report

For students (instructor review required)

ⓒ2019 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| Fake News Classification |

17-10-2022

Aces

Hrushikesh

Joseph Srujan

Roshan

Devendra

Sumanth

Naveen

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. IoT Service Model

2.2. Data Processing

2.3. Service Implementation

2.4. System Diagram

3. Results

3.1. Data Acquisition (Sensor, Actuator, Controller)

3.2. Network & Communication

3.3. Hardware Implementation

3.4. Data Visualization

3.5. Testing and Improvements

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

**1.1. Background Information**

A few decades ago, the term "Fake News" was less obscure and less common,

but in the digital age of social media, it has emerged as a big monster. Our

culture is facing rising issues with fake news, information bubbles, news

manipulation, and a loss of faith.

In the media. However, in order to begin tackling this issue, a thorough

knowledge of fake news and its sources is necessary. Only after that can one

consider the many methods, such as artificial intelligence (AI) and natural

language processing (NLP), that can aid in combating this issue. It may happen so rapidly that measuring fake news or even accurately describing it would no

longer be an objective statistic. In its purest form, fake news is entirely untrue and has been edited to look like reliable reporting, garner as much attention as possible, and increase advertising income. Despite all of these flaws, some organizations have made an effort to classify fake news in various ways.

Some of unknown terms are

1.LSTM (Long - term short memory)

2.CNN

3.Word cloud

4.Tokenization

5.Word2Vector

6.Padding sequences

**1.2 Motivation and Objective**

The widespread problem of fake news is very difficult to tackle in today’s

digital world where there are thousands of information sharing platforms

through which fake news or misinformation may propagate. It has become a

greater issue because of the advancements in AI which brings along artificial

bots that may be used to create and spread fake news. The situation is dire

because many people believe anything they read on the internet and the ones

who are amateur or are new to the digital technology may be easily fooled. A

similar problem is fraud that may happen due to spam or malicious emails and messages. So, it is compelling enough acknowledge this problem take on this

challenge to control the rates of crime, political unrest, grief, and thwart the

attempts of spreading fake news.

**1.3 Members and Role Assignments**

Hrushikesh,Joseph Srujan: Collected the datasets and checking which model will be suitable for our project and knowing the models thoroughly.

Devendra,Sumanth : Worked with several data visualization

techniques and several data pre-processing

techniques.

Roshan,Naveen :Worked with different datasets and helped

to get high accuracy LSTM model.

**1.4 Schedule and Milestones**

2/10/2022 - 6/10/2022: Learnt about useful libraries and modules to implement

a new mode and collected the data sets from Kaggle

and git-hub.

7/10/2022 - 8/10/2022: Performed different data visualization and data

preprocessing techniques which are required for

dataset.

9/10/2022 - 11/10/2022: Understanding and implementing the LSTM

technique on two different data sets.

12/01/2022 - 13/10/2022: Implementing the CNN model for the data set.

2. Project Execution

**2.1 Data Acquisition**

We have taken datasets from Kaggle and GitHub. In this news articles and Us

news, it consists of Separate files like real and fake. First, we will do operations on fake news which consists of columns, which consists of title, text, subject,

date. Here we need only the text and the title. we first Change the text into

sequences and then we pad up them. we convert that to vectors and will trained by using CNN and LSTM model.

**2.2 Training Methodology**

We have taken two datasets (.csv) files one from Kaggle and other from GitHub sources. The two datasets consist of data related to Us News but with different

features. It will be labelled as real and fake to train the model. A Deep learning model cannot be made by giving inputs as words, we tokenize the words into

vectors. We will be Manipulate the data by giving the input as vectors and then performs as sequential operation and then it uses the pad sequences to get the

vector of same length. Then we give the data for the training.

**2.3 Workflow**

Imported the required libraries, modules and packages.

Dataset has been taken from the Kaggle and GitHub for test and train data

Visualizing the text which is in the dataset with wordcloud and countplot

Assigning the labels to y and total text to x and splitting the data to test and train.

Tokenizing the text into words and covert into sequences.

We pad up the sequences in order to get the same length of the array.

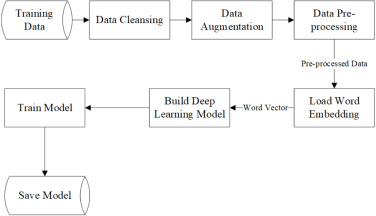
We have used sequential model for the sequences after padding.

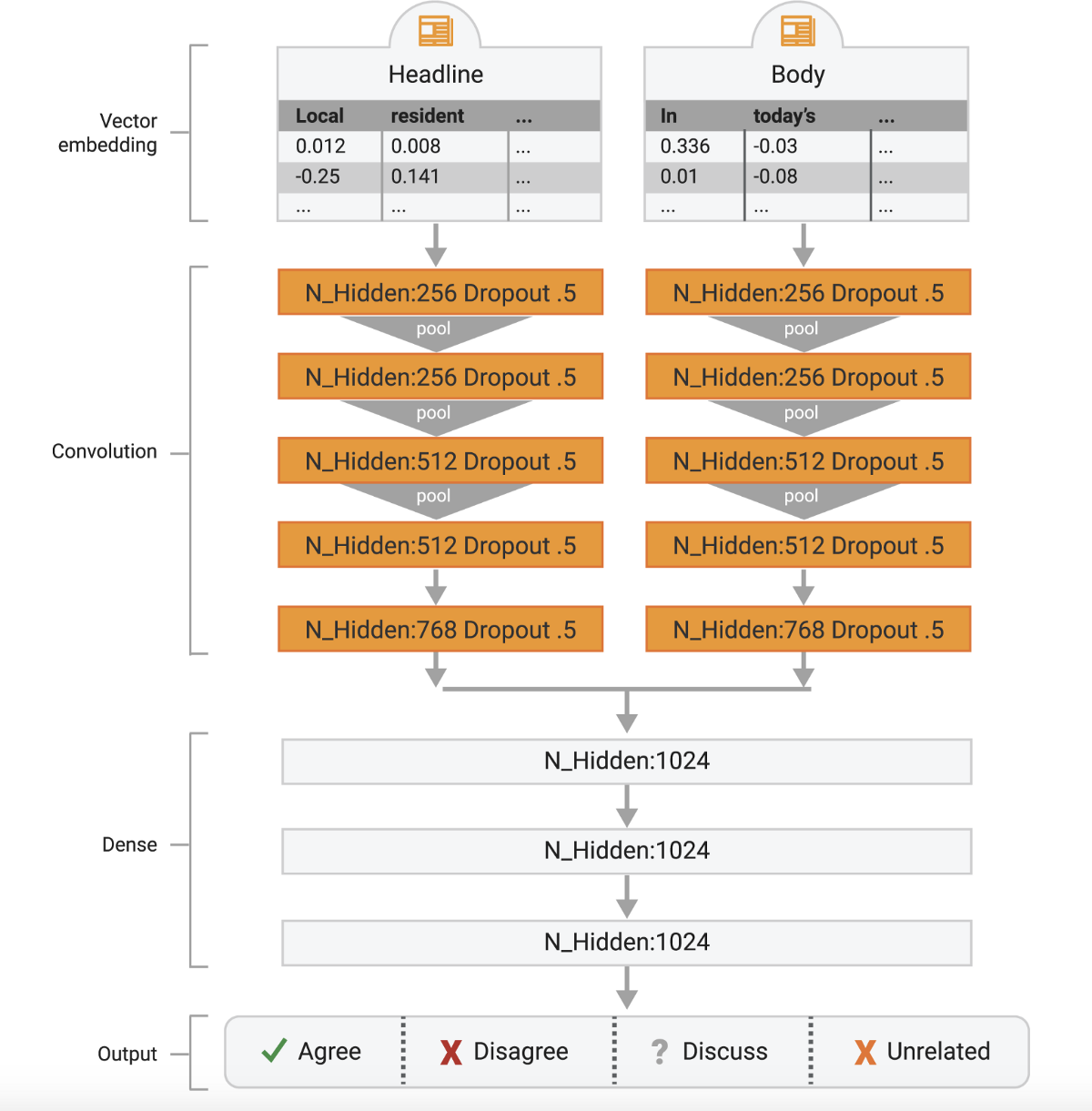
In Sequential model we have added embedding and LSTM/CNN layer for our

model

We will be predicting accuracy and confusion matrix for our model.

**2.4 System diagram**





3. Results

**3.1. Data Preprocessing**

Data pre-processing is utilized to represent complex structures with attributes,

binarize attributes, change discrete attributes, persist, and manage lost and

obscure attributes. During data pre-processing, different visualization procedures are helpful. A cautious pre-processing strategy is required to ingest the data in a neural network for fake news detection because social media data sources are

fragmented, unstructured, and noisy.

It is a popular fact that amid the learning stage, data pre-processing saves

computational time and space. In addition, limiting the impact of artifacts

during the learning process, text pre-processing avoids every ingests of noisy

data. The data becomes a logical representation after proper text pre-processing. It also included the most representative descriptive words. Experimented on a

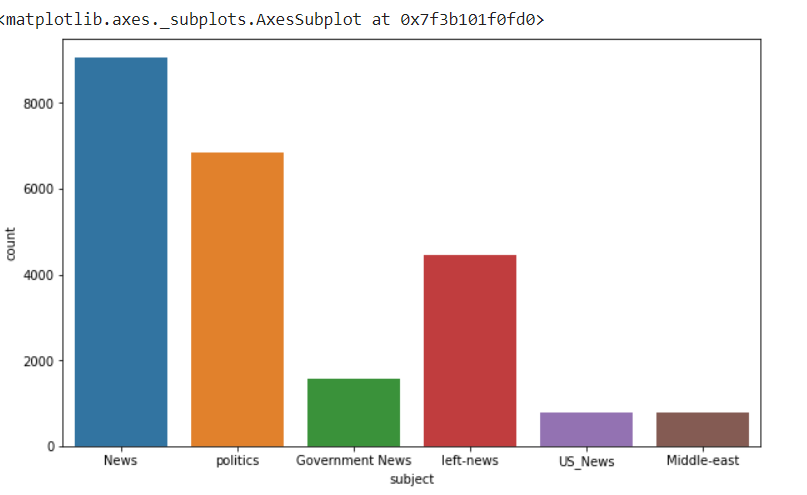
fake news detection model in which the accuracy was only 78% when they used the features excluding data cleaning or pre-processing, which is surprisingly

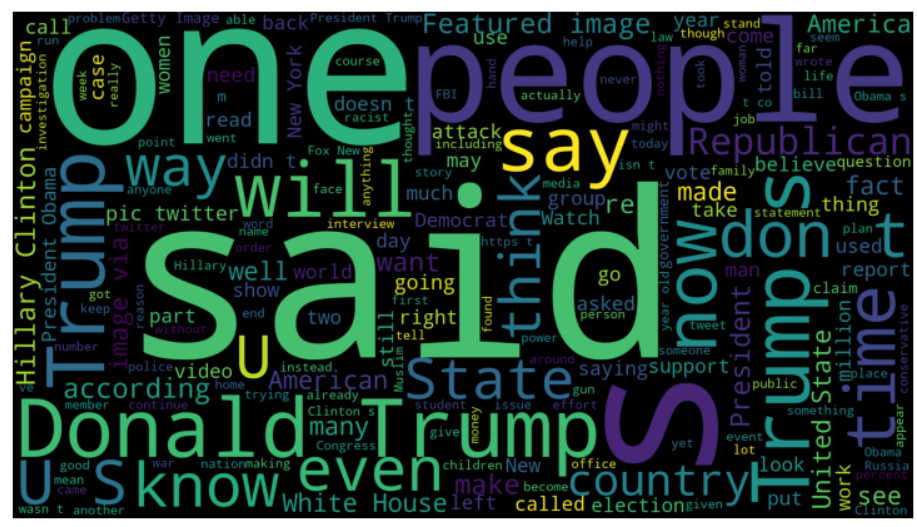
poor. After performing the pre-processing steps and removing unnecessary data, the accuracy increases dramatically to 99% Data quality assessment,

dimensionality reduction, and splitting of the dataset are the data pre-processing steps used in various studies.

**3.2 Exploratory Data Analysis (EDA)**

We have used several data visualization techniques before the data preprocessing i.e., Word cloud and bar plot in order to know the word frequency and know the relation between numerical variables and categorical variables.





**3.3 Modeling**

**CNN:**

A few deep learning models have been introduced to handle ambiguous

detection issues. CNNs and RNNs are the most interesting models.

Researchers are trying to boost the performance of the fake news

detector with CNN by taking its power of extracting features well and

better classification process.

However, CNNs are also gaining popularity in the NLP technique too. It is

utilized for mapping the features of n-gram patterns. The CNN is similar to a

Multilayer perceptron (MLP) as it is an unsupervised multilayer feed-forward

neural network. The CNN consists of an input layer, an output layer,

and a sequence of hidden layers. CNNs are mostly used for picture recognition and classification.

**LSTM:**

Long Short-Term Memory networks (LSTM) are a special type of RNN

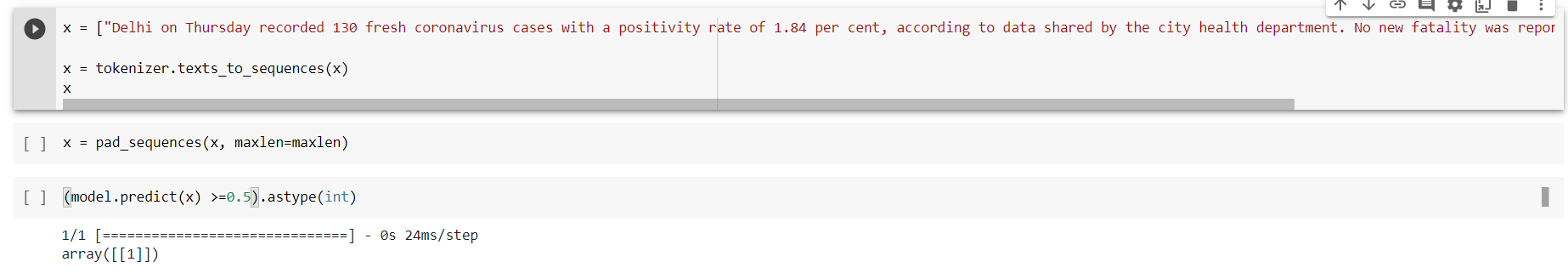
competent in learning long-term dependencies. LSTM is a very effective

solution for addressing the vanishing gradient problem. In LSTM-RNN the

hidden layer of basic RNN is replaced by an LSTM cell.

**3.4 User Interface (Interface)**

Here, after training the dataset with the model. We will be giving text as input and this text is converted into sequences. Then we apply our model in order to classify real or fake.



**3.5. Testing and Improvements.**

Through training the model over several epochs, the weights of the trained

model are updated. To create the final network, the weights and biases of the

network are trained over the epochs.To improve the final accuracy we use

epochs. We have done CNN and LSTM models, out of which LSTM has got

highest accuracy. So we will be using LSTM for our model.

4. Projected Impact

**4.1. Accomplishments and Benefits**

Since the dawn of human society, there has always been false information. But

as the global media environment changes and modern technology advance, the

circulation of false information grows. Fake news may have significant negative effects on the social, political, and economic contexts. Fraudulent news and

fake information come in many forms. Fake news has a significant effect

because information shapes our perspective of the world. On the basis of the

information, we make important decisions. We build an opinion about a

circumstance or a group of people based on the information we learn. If we find manufactured, skewed, or incorrect information online, we cannot make wise

selections. The following are the main effects of fake news:

**Impact on innocent people :** Rumors can have a significant impact on a

number of innocent people. Social media may be used to harass these individuals. Additionally, they could come across insults and threats that could have

far-reaching effects. People shouldn't make snap judgments about others or

accept false information posted on social media.

**Impact on Health:** An rising number of people are using the Internet to find

news about their health. The lives of people could be affected by fake news in

the health field. Therefore, one of today's biggest problems is this. In the past

year, false information regarding health has had a significant influence. As a

result of pressure from physicians, legislators, and health advocates, social

media companies have changed several of their policies to prohibit or restrict

the dissemination of health misinformation.

**Financial Impact**: In the business sector and other industries, fake news is

currently a major issue. In order to increase their earnings, dishonest business

people circulated false information. Stock prices may drop as a result of false information. It can destroy a company's reputation. Customers' expectations are

impacted by fake news as well. Fake news has the potential to breed

unscrupulous commercial practices.

**4.2 Future Improvements**

Finally, we want to extend this work by performing similar analysis on a

completely different dataset such as Twitter and Facebook. By classifying fake

news from social media platforms, we hope to get one step closer towards

building an automated fake news detection platform. This study provides a

baseline for the future tests and broadens scope of the solutions dealing with

fake news detection. The social media data will ensure that the variations in the language are taken care of. We would like to further dig deep and evaluate the

effects of such news propagation on the readers and come up with simple

techniques for faster prediction. The research can borrow qualitative models

built on similar tasks by other disciplines and re-evaluate feature engineering

and preprocessing techniques used.

5. Team Member Review and Comment

|  |
| --- |
| IMG_20221020_111415 |

|  |  |
| --- | --- |
| NAME | REVIEW and COMMENT |
| HRUSHIKESH |  |
| JOSEPH SRUJAN |  |
| DEVENDRA |  |
| SUMANTH |  |
| NAVEEN |  |
| ROSHAN |  |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/20 |  |
| CODING | \_\_/20 |  |
| PROJECT MANAGEMENT | \_\_/30 |  |
| PRESENTATION & REPORT | \_\_/30 |  |
| TOTAL | \_\_/100 |  |