**DISASTER TWEETS DETECTION IN TWITTER DATABASE**

Project Report

Submitted in the partial fulfillment of the requirements for the award of the degree of

# Bachelor of Technology

# In

Department of Computer Science Engineering

By

A.SAHITI SANDYA VANI (170030056)

M.B.KRISHNA VENI (170030759)

M.SAI SPANDHANA (170030786)

CH.H.V.L. NARSIMHA RAO (170031500)

under the supervision of

# **Dr M.SREE DEVI**

**PROFESSOR**



Department of CSE

K L E F, Green Fields, Vaddeswaram- 522502, Guntur(Dist), Andhra Pradesh, India. december, 2020



# K L E F deemed to be UNIVERSITY

**Green fields, Vaddeswaram, Guntur Dist**.

**DECLARATION**

The project Report titled “DISASTER TWEETS DETECTION IN TWITTER DATABASE” is a record of genuine work of A. Sahiti Sandhya Vani (170030056), M. B. Krishna Veni (170030759), M. Sai Spandhana (170030786), CH. H.V. L.Narsimha Rao (170031500) submitted in partial fulfillment for the award of B.Tech in Computer Science Engineering to the K L University. The results presented in our report have not been copied from any other departments /Universities/ Institutions.

A.SAHITI SANDHYA VANI (170030056)

M.B.KRISHNA VENI (170030759)

M.SAI SPANDHANA (170030786)

CH.H.V.L.NARSIMHA RAO(170031500)



# K L E F deemed to be UNIVERSITY

**Green fields, Vaddeswaram, Guntur Dist**.

**CERTIFICATE**

This is to certify that the Project Report entitled “DISASTER TWEETS DETECTION IN TWITTER DATABASE” is being submitted by A. Sahiti Sandhya Vani (170030056) , M. B. Krishna Veni (170030759), M. Sai Spandhana (170030786) , CH. H. V. L. Narasimha Rao (170031500) in partial fulfillment for the award of B.Tech in Computer Science Engineering to the K L University is a record of genuine work carried out under our guidance and supervision.

The results presented in our report have not been copied from any other departments /Universities/ Institutions.

## Signature of the Supervisor

Dr.M.SREEDEVI

## Signature of the HOD Signature of the External Examiner



**K L E F deemed to be UNIVERSITY**

**Green fields, Vaddeswaram, Guntur Dist**.

## ACKNOWLEDGEMENT

We express our sincere gratitude to **Dr. M. Sreedevi** madam for encouraging and guiding us to undertake this project work. We express our deep sense of gratitude to our beloved course lecturers of department for their encouragement.

Place: Vaddeswaram

Date:

|  |  |
| --- | --- |
| Name | Id.no |
| **A.SAHITI SANDYA VANI** | **170030056** |
| **M. B. KRISHNA VENI** | **170030759** |
| **M. SAI SPANDHANA** | **170030786** |
| **CH.H.V.L.NARSIMHA RAO** | **170031500** |

**ABSTRACT**

Fake news is inferior quality news with purposefully false information to believe it. Social networking sites that attract millions of users around the globe. Fake news on social media is making an appearance that is attracting a huge attention. Fake news is purposefully composed to fake information. This kind of situation could bring a great conflict in real time. Our project explores the data mining techniques to predict whether given news is real or fake and clean the data. After this process our detection model got created by using five different techniques. And during predicting the result we found something interesting that if we use an image in the place of dataset it will be easy to predict and helps us if we are having any trouble in uploading the large dataset by using LSTM.

.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| S.no | Title | Pageno |
| 1 | Introduction | **7** |
| 2 | Literature Survey | **8** |
| 3 | Theoretical Analysis | **13** |
| 4 | Experimental Investigations | **16** |
| 5 | Experimental Results | **23** |
| 6 | Discussion of Results | **30** |
| 7 | Conclusion | **31** |
| 8 | References | **32** |

**INTRODUCTION**

These days, many of the people give priority to read the news through social media. In recent years, the popularity of social media has increased significantly. There are millions of news contents published on the internet community replaced traditional community and became the main platform in terms of spreading news.

Wherever, Twitter App is not only based on individual opinions, but also for official information of a particular occasion can be delivered, as an organization can use Twitter like an individual user.

Users who have interest in a certain area such as medical treatment, leisure, or any interesting hobby can gather and conveniently post informative contents. Twitter can likewise be utilized in a different term, and it's worth might be boundless later.

Knight Foundation started a study to investigate how fake news spread on the platform before, during and after the 2016 U.S. presidential election.

The examination analysed more than 10 million tweets from 700,000 Twitter accounts that linked to more than 600 fake and conspiracy news sites. It identified clusters of Twitter accounts that linked back to these sites repeatedly, often in ways that seemed coordinated or even automated.

Fake news is a false or a misleading information presented as news. It often has an aim of damaging the reputation of a person. First, fake news is 8 intentionally written to mislead consumers, it does not make satisfactory to spot fake news from news itself.

You can report directly from an individual Tweet, List, or profile for certain violations, including spam, abusive or harmful content, inappropriate ads, self-harm and impersonation. For information about reporting other types of violations but even to report the spam issues or fake tweets we need to identify the fake tweets so to predict such fake information we are about to use 5 algorithms and it would be helpful to report all kinds of harmful rumours.

**LITERATURE SURVEY**

**2.1.1 Title:**

A Study on Fake News Detection Using Naïve Bayes, SVM (Support Vector Machine), Neural Networks and LSTM.

**2.1.2 Proposed by:**

Diana Elizabeth Roy, Prannay S Reddy

**2.1.3 Introduction:**

This Paper is composed by Prannay S Reddy he dealt with discovering how news can be separated as evident or not by characterizing on a couple of articles that are consistently experienced in phony news. the attributes depend on "social speculations and brain science, existing calculations from an information mining point of view and agent datasets." it likewise investigations the different difficulties one will experience while contemplating this subject. He proposed an arrangement technique which can be utilized continuously information that gets made on Facebook. It can likewise be utilized in other online media like Twitter and WhatsApp. He has used a straightforward Naive Bayes arrangement to characterize the approaching information as dependable or non-reliable.

**2.1.4 Methodology:**

The system in this paper depends on a schematic portrayal of the order cycle. The initial phase in cycle is to pre-measure the information in the dataset which finishes up stop lower packaging and word evacuation of the whole characters. Additionally, extraordinary characters are taken out from the corpus. The Doc2Vec is utilized, and a vector is shaped speaking to the reports in question. The urgent advance of the classifier is to anticipate the class of the given news story into genuine or counterfeit. There are of 4 AI models being tried here to show up at the most ideal choice. The information is isolated into train and test. The preparation information has 7000 corpus of news stories and the testing information has 26000 corpus of news story.

The other technique in this paper is LSTM, which represents Long momentary memory is an expansion of the ongoing acclaimed RNN (Recurrent Neural Network). Notwithstanding RNN, LSTM's additionally has memory as time goes on. It includes three doors in particular info, entryway, yield entryway and fail to remember door.

**2.2.1 Title:**

Sentiment analysis on twitter data

**2.2.2 Proposed by:**

Onam Bharti,Mrs. Monika Malhotra

**2.2.3 Introduction:**

In this paper they described about hints to do opinion mining for unstructured data on web. At present, opinion mining concentrates for subjective texts and search objective statements which carry opinions. So, they described innovative approach to define and control subjective as well as objective texts for opinion mining.

**2.2.4 Methodology:**

The suggested design of four modules are Feature Clustering using Modified K-means, user interface, log pre-processing, Naïve Bayes Classification testing and training using k-nearest neighbors algorithm for best Précised classification of viewpoint. The design could find unrelated data and more accuracy using K means Modified with Naïve Bayes algorithm. Present research is searching on Sentimental analysis regarding complete volume of opinion of rich web sources such as communication groups like forums, feedbacks are in digital form. Many ways that got improved for main problems of opinion mining's and got solved. We received a picture of which are involved in making a software system for sentimental analysis regarding the analysis.

**2.3.1 Title:**

Techniques to Detect Spammers in Twitter- A Survey

**2.3.2 Proposed by:**

Monika Verma, Sanjeev Sofat

**2.3.3 Introduction:**

In this paper we focused on collecting the issues based on Online Social Networking sites which are capable to secure and privacy issues because of the quantity of user information being processed by these sites each day. Users of social sites are shown to many attacks. Combined with the casual problems like spamming, phishing attacks, malware infections, social bots. The Best challenge that social networking sites present for users is to keep confidential data secure and confidential.

**2.3.4 Methodology:**

There are various techniques have been used by researchers to find out the fraud profiles in various OSNs. We are focusing only on the work that have been done to identify fraud identities in Twitter as it is not only a social communication media but in fact is used to share and spread information related to trending topics in real time like K-pop idol events like BTS(Bangtan Sonyeondan)

This survey of existing methods for detecting spam profiles in OSNs has been done after a systematic review with principled approach in which major research databases for Computer Science have been searched like Institute of Electrical and Electronics Engineers Xplore, Atmospheric Chemistry & Modelling Digital Library, SpringerLink, Google Scholar, ScienceDirect for concerned topic.

**2.4.1 Title:**

Fake news detection in social media

**2.4.2 Proposed by:**

Kelly Stahl

**2.4.3 Introduction:**

In this paper the Quick technological advancement has newspapers and journalism to be exposed over the web and the rise of Twitter and some other social networking sites. Social Sites have become a noteworthy method to speak for people with each other and offer schemes and thoughts. Vast components of a person these networking sites is quick sharing of information. The only communication of sharing is WebSourced these days so this effect the real and reel.

As such, fake news impacts have spread deliberately in the past and something must be done to keep this from proceeding later in future. This project includes using AI, ML and NLP techniques to make a model that can uncover records, with high chance of fake news stories and articles. A Substantial number of the current computerized ways to deal with this issue were based on a "boycott" of creators and sources that were known makers of fake news.

**2.4.4 Methodology:**

The methods in this paper are the techniques to make a model that will uncover records with high probability rate of false tweets. AI calculations however not restricted to convolutional neural systems, recurrent neural systems. The result of this project should be to decide how much can be accomplished in this task by dissecting designs contained in the text and bind to the outside data about the world. This kind of answer is not expected to be an end-to-end solution for fake tweets. Like the "boycott" approaches referenced, there are cases in which it fails and some for which it succeeds. Rather than being an end-to-end solution, this method is expected to be one solution that could be utilized to help people who are attempting to differ fake news. Then again, it may be one technique which could be utilized in coming applications that bravely combine different devices to process an end-to-end solution for robotization of cycle of phony news order

**2.5.1 Title:**

Identifying Fake News and Fake Users on Twitter

**2.5.2 Proposed by:**

Costel-Sergiu Atodiresei\*, Alexandru Tănăselea, Adrian Iftene

**2.5.3 Introduction:**

In this paper we extracted about the general architecture of the system. There was a Twitter crawler component, which collects tweets and adds them to our database. When we will need tweets from trustworthy resources to compare with our present one, we can retrieve them directly from our database.

The Processing module: when a user wants to know the credibility of a new tweet, he inputs the link of the tweet in our interface. This algorithm then uses a Named Entity Recognition component, which separates the text into its composing parts: it brings out the entities (nouns and their relative importance in the context), the topics, the social tags, the overall tweet sentiment, and the hashtag sentiment.

**2.5.4 Methodology:**

The algorithm analyzes the link received and returns a string message that includes the Decision, the User score, and the Tweet Score. The twitter link is included of twitter address, user screen name and status id (tweet id). Twitter has a user-friendly manner of taken this tweet link.

Next, algorithm works as it follows:

• It receives the user details from our database.

• If we do not have the user in the database, it gets its values from Twitter and initialize its score to zero.

• It gets the tweet’s details from Twitter;

• Creates a tweet score code based on the received data from different API’s

• Creates a user score code as an arithmetic mean of its prior score and actual tweet score.

**THEORETICAL ANALYSIS**

**3.1 PURPOSE:**

To predict the given news is real or fake news, we will train classification algorithm. And we implied Long short-term memory [LSTM] and predicted the accuracy of disaster tweets in twitter database.

**3.2 DATA PREPROCESSING:**

Data preprocessing is one of the data mining technique which converts the raw data into understandable, useful and efficient format. Real world data is incomplete, inconsistent and it contains many errors which affects the results. So, all these issues can be resolved by performing data preprocessing.

**3.2.1 DATA CLEANING :**

Data cleaning is the process of removing incorrect, duplicate or incomplete data within a dataset. When we combine multiple data sources there are many chances for data to be duplicated.

**BENEFITS OF DATA CLEANING:**

1) having clean Data will increase overall productivity.

2) It removes major errors and inconsistencies when multiple sources of data are pulled into one dataset.

**3.2.2 DATA INTEGRATION:**

Data Integration is a data preprocessing technique that combines data from multiple different sources into a reasonable data and provide unified view of the data

**BENEFITS OF DATA INTEGRATION:**

1) simple and speedy relations

2) combine data from various origins

3) usage of data

4) greater cooperation

5) quality and integrity of data

6) improves data availability.

**3.2.3 DATA TRANSFORMATION:**

Data transformation is a data preprocessing technique that converts source system into the required format of a destination system. It is the process of changing the format, structure or values of data.

**BENEFITS OF DATA TRANSFORMATION:**

1) Data is transformed to make it better organized.

2) The transformed data is easier for both humans and computers to use.

3) Properly formatted and validated data improves data quality.

**3.2.4 DATA REDUCTION:**

Data reduction is a data preprocessing technique that reduces the amount of capacity required to store the data. It can increase storage efficiency and reduce the costs.

**BENEFITS OF DATA REDUCTION:**

1) smaller disk space.

2) speedy transfer of files, as it contains smaller disk space it reduces chances of disk collapse.

3) it contains more capacity.

**3.3 METHODOLOGY:**

**3.3.1 LONG SHORT-TERM MEMORY**

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. This is a behavior required in complex problem domains like machine translation, image & speech recognition. LSTMs are a complex area of deep learning.

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series.

It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic.

**ADVANTAGES:**

LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an **advantage of LSTM** over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

**DISADVANTAGES:**

In LSTM it is in-appropriate to imply the dropout algorithm where in d-algorithm the input units are excluded from the activation.

**4: EXPERIMENTAL INVESTIGATION**

**4.1 CODE:**

import tensorflow as tf

from tensorflow import keras

import pandas as pd

import numpy as np

import os

import matplotlib.pyplot as plt

import time

df = pd.read\_csv("train.csv")

df

df.shape

df.head()

print((df.target == 1).sum()) #disaster

print((df.target == 0).sum()) #no disaster

import re

import string

def remove\_URL(text):

url = re.compile(r"https?://\S+|www\ .\S+")

return url.sub(r"", text)

def remove\_punct(text):

translator = str.maketrans("","",string.punctuation)

return text.translate(translator)

string.punctuation

pattern = re.compile(r"https?://(\S+|www)\.\S+")

for t in df.text:

matches = pattern.findall(t)

for match in matches:

print(t)

print(match)

print(pattern.sub(r"", t))

if len(matches) > 0 :

break

df["text"] = df.text.map(remove\_URL)

df["text"] = df.text.map(remove\_punct)

#remove stopwords

#pip install nltk

import nltk

nltk.download('stopwords')

from nltk.corpus import stopwords

#Stop Words: A stop word is a commonly used word (such as "the", "a" , "an" , "in") that a search engine

#has been programmed to ignore, both when indexing entries for searching and when retrieving them

#as the result of a search query.

stop = set(stopwords.words("english"))

def remove\_stopwords(text):

filtered\_words = [word.lower() for word in text.split() if word.lower() not in stop]

return " ".join(filtered\_words)

stop

df["text"] = df.text.map(remove\_stopwords)

df.text

from collections import Counter

#count unique words

def counter\_word(text\_col) :

count = Counter( )

for text in text\_col.values :

for word in text.split( ):

count[word] += 1

return count

counter = counter\_word(df.text)

len(counter)

counter

counter.most\_common(5)

num\_unique\_words = len(counter)

#split dataset into training and validation set

train\_size = int(df.shape[0] \* 0.8)

train\_df = df[ :train\_size]

val\_df = df[train\_size: ]

#split text and labels

train\_sentences = train\_df .text. to\_numpy( )

train\_labels = train\_df.target.to\_numpy( )

val\_sentences = val\_df.text.to\_numpy( )

val\_labels = val\_df.target.to\_numpy( )

train\_sentences.shape, val\_sentences.shape

#tokenize

from tensorflow.keras.preprocessing.text import Tokenizer

#vectorize a text corpus by turning each text into a sequence of integers

tokenizer = Tokenizer(num\_words=num\_unique\_words)

tokenizer.fit\_on\_texts(train\_sentences) #fit only to training

#each word has unique index

word\_index = tokenizer.word\_index

word\_index

train\_sequences = tokenizer.texts\_to\_sequences(train\_sentences)

val\_sequences = tokenizer.texts\_to\_sequences(val\_sentences)

print(train\_sentences[10:15])

print(train\_sequences[10:15])

#pad the sequences to have the same length

from tensorflow.keras.preprocessing.sequence import pad\_sequences

# Max number of words in a sequence

max\_length = 20

train\_padded = pad\_sequences(train\_sequences, maxlen=max\_length, padding="post", truncating="post")

val\_padded = pad\_sequences(val\_sequences, maxlen=max\_length, padding="post", truncating="post")

train\_padded.shape, val\_padded.shape

train\_padded[10]

print(train\_sentences[10])

print(train\_sequences[10])

print(train\_padded[10])

#check reversing the indices

#flip (key,value)

reverse\_word\_index = dict([(idx, word) for (word, idx) in word\_index.items()])

reverse\_word\_index

def decode(sequence):

return " ".join([reverse\_word\_index.get(idx, "?") for idx in sequence])

decoded\_text = decode(train\_sequences[10])

print(train\_sequences[10])

print(decoded\_text)

#create LSTM model

from tensorflow.keras import layers

model = keras.models.Sequential()

model.add(layers.Embedding(num\_unique\_words, 32, input\_length=max\_length))

model.add(layers.LSTM(64, dropout=0.1))

model.add(layers.Dense(1, activation="sigmoid"))

model.summary()

loss = keras.losses.BinaryCrossentropy(from\_logits=False)

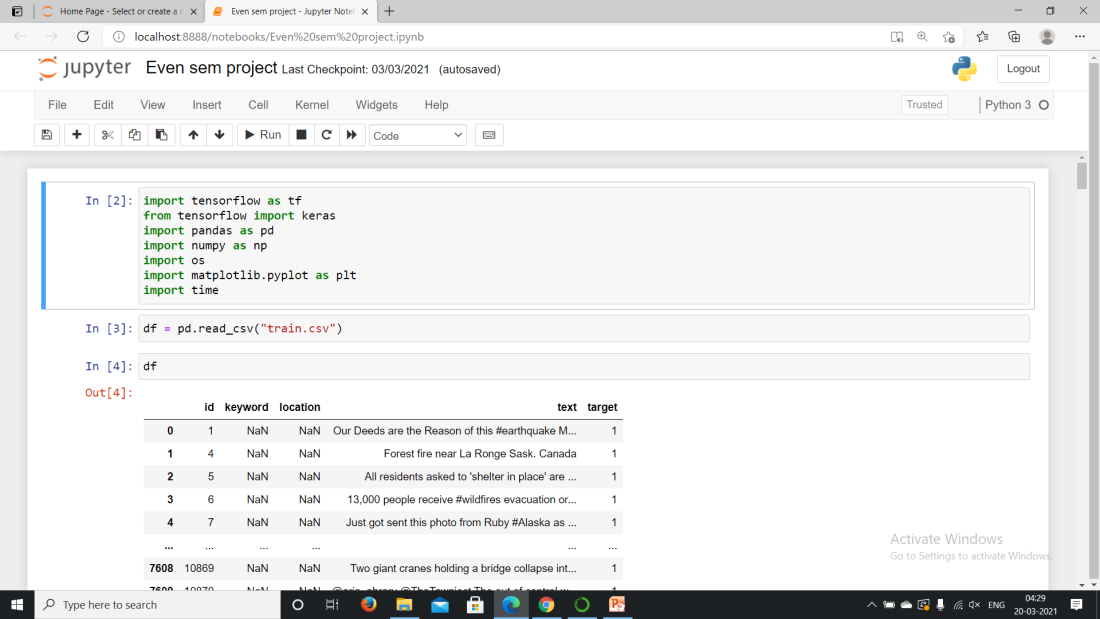
optim = keras.optimizers.Adam(lr=0.001)

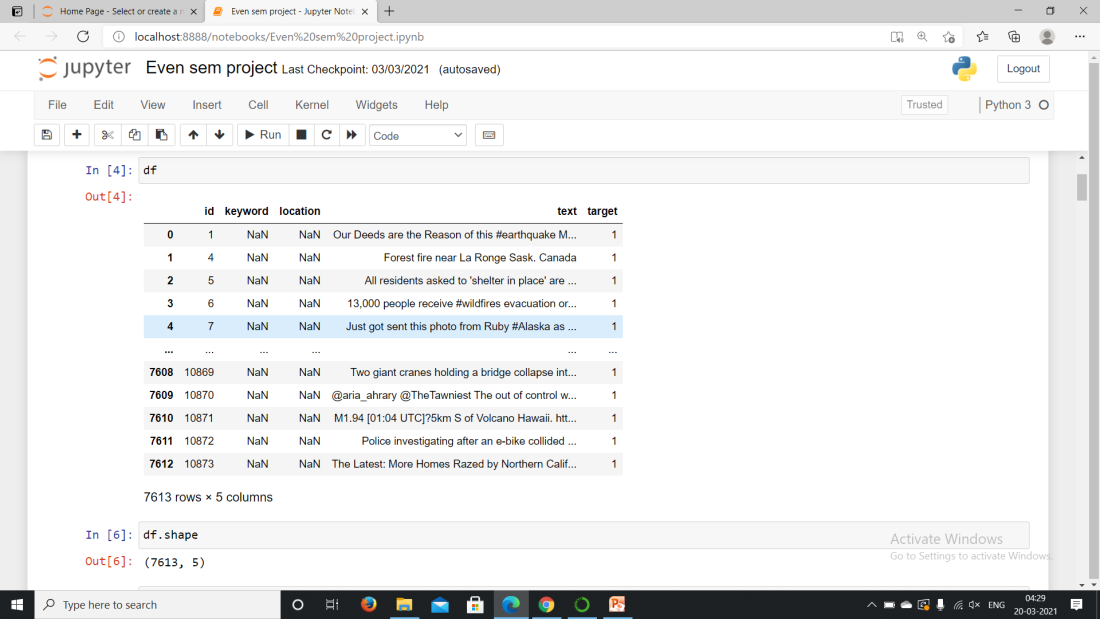
metrics = ["accuracy"]

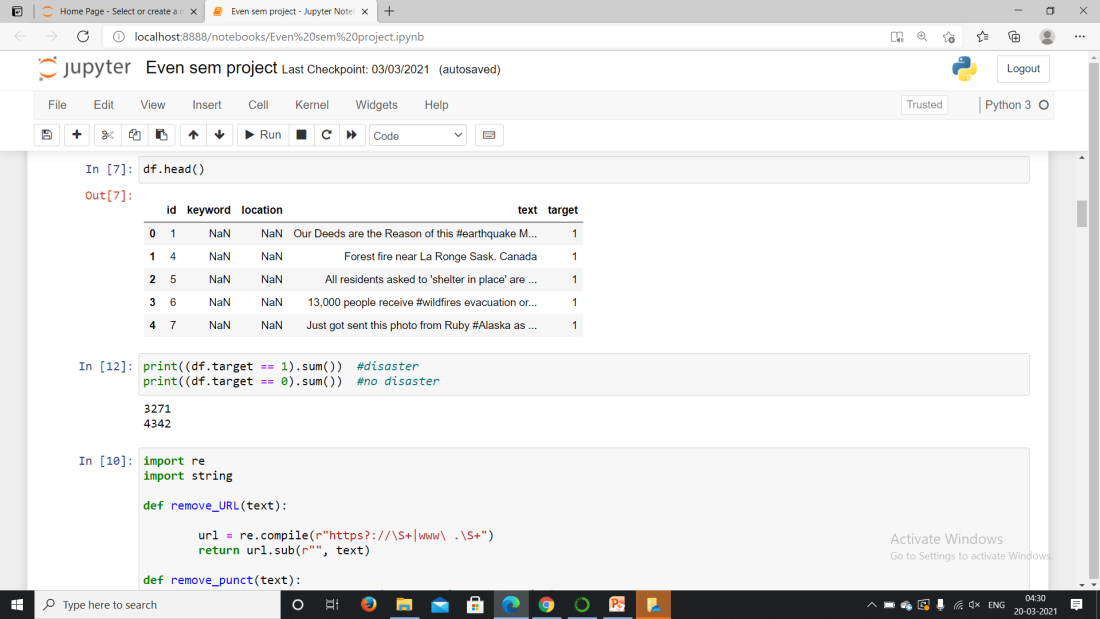
model.compile(loss=loss, optimizer=optim, metrics=metrics)

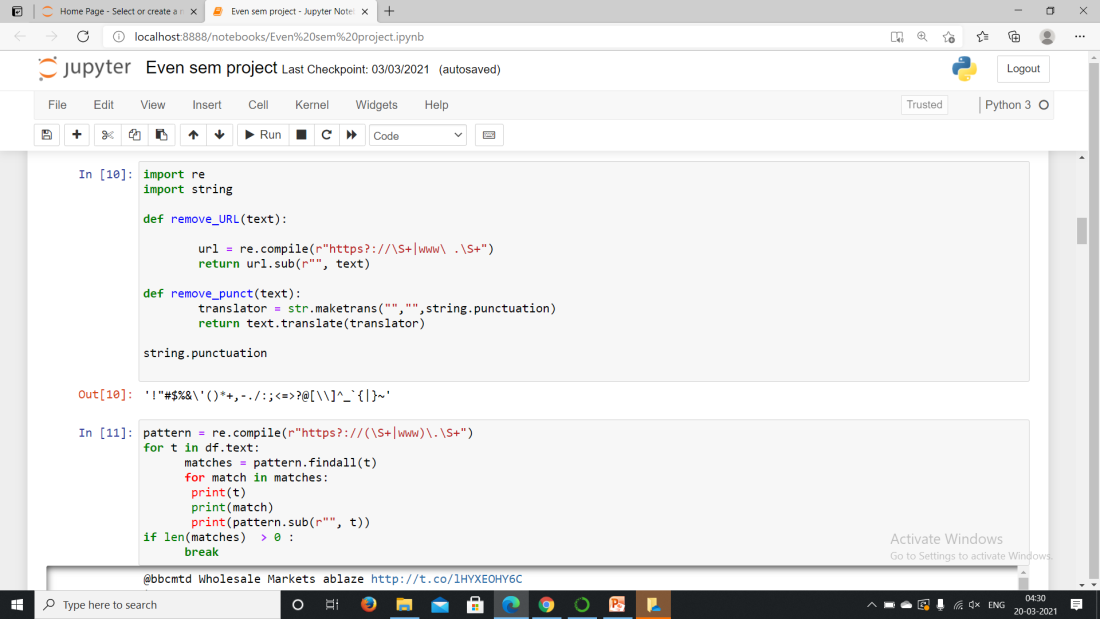
model.fit(train\_padded, train\_labels, epochs=20, validation\_data=(val\_padded, val\_labels), verbose=2)

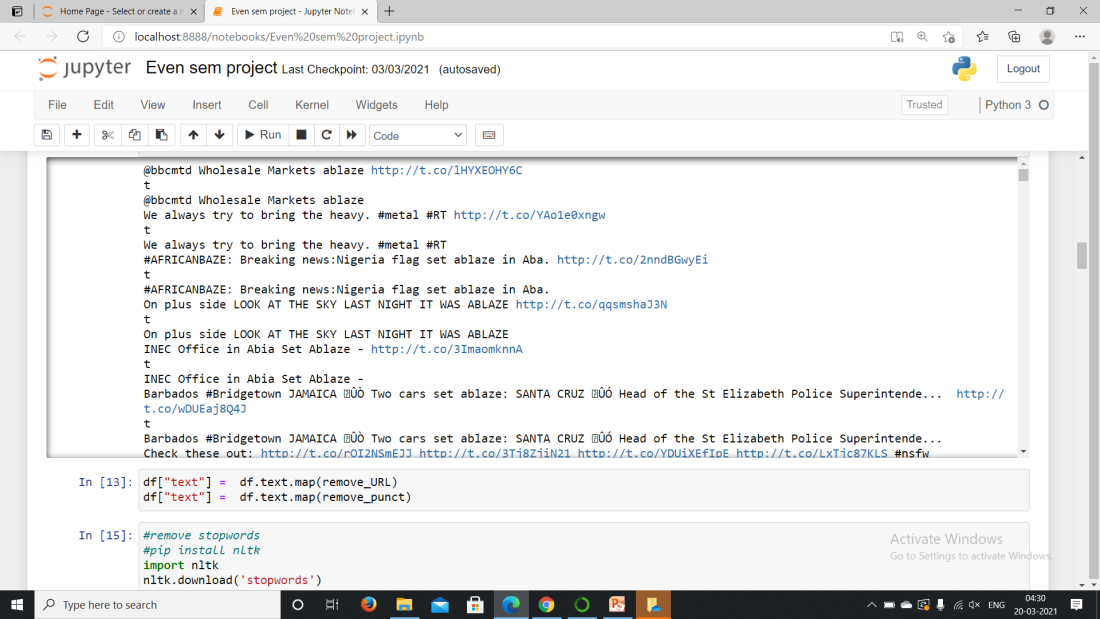
**5: EXPERIMENTAL RESULT**

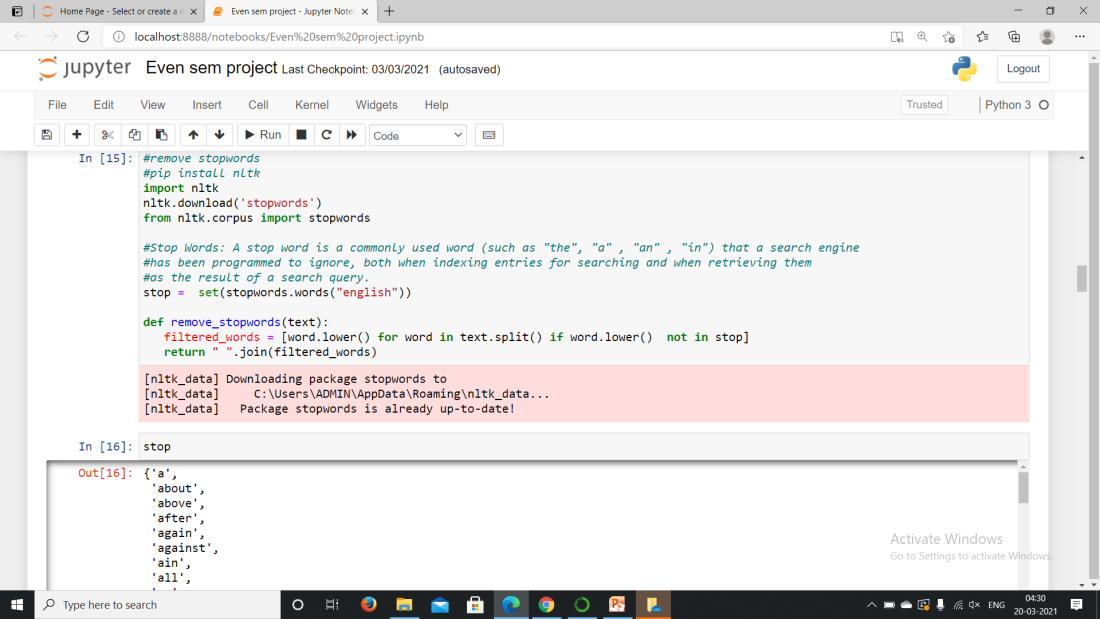


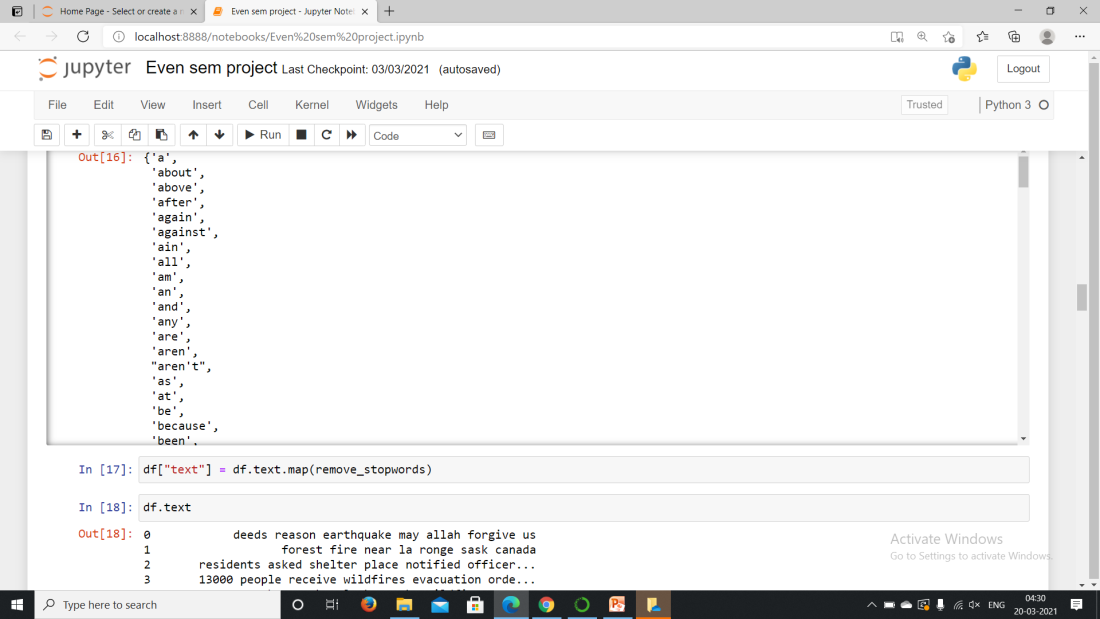


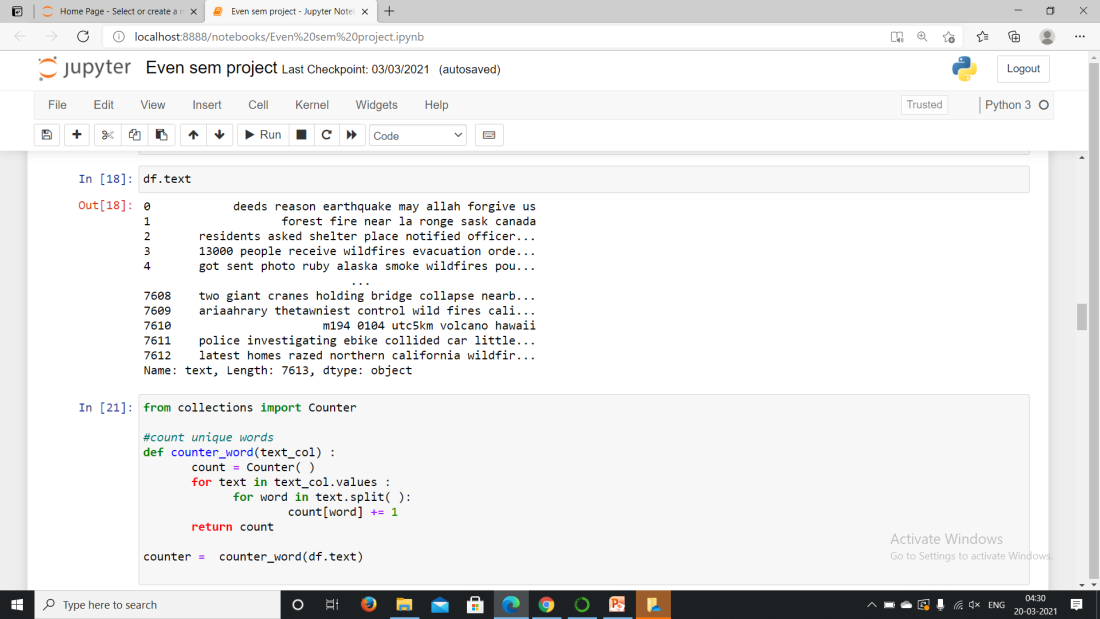


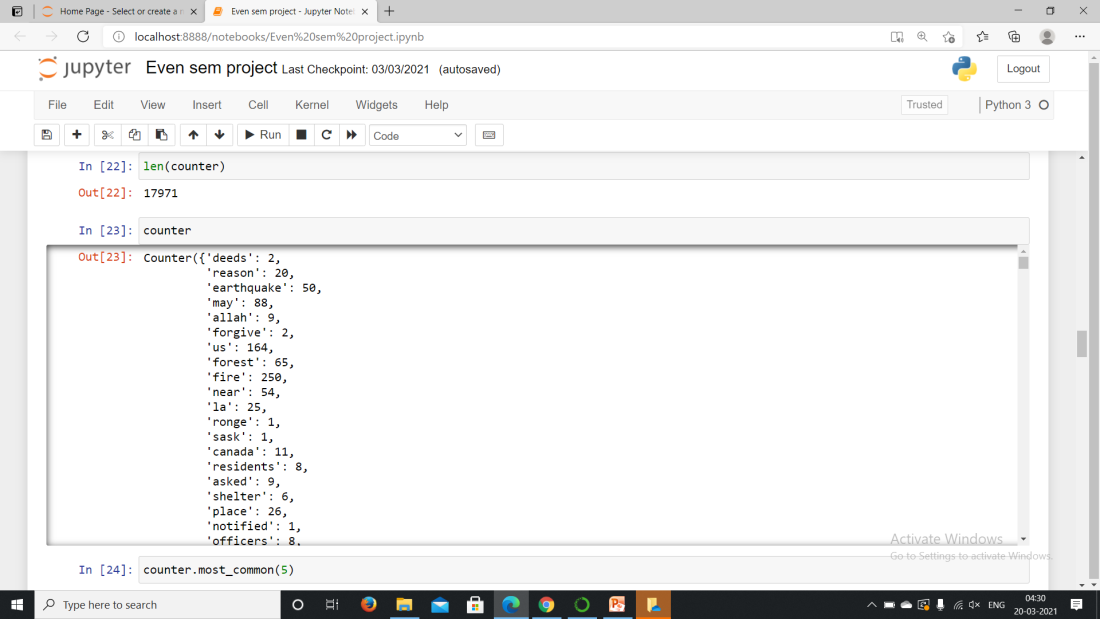


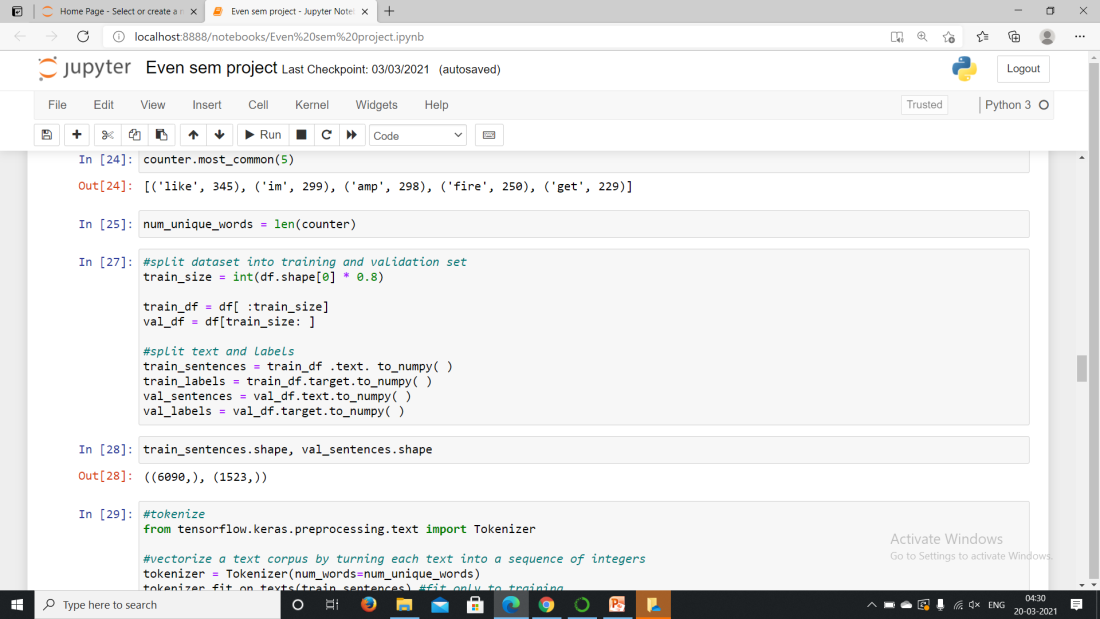


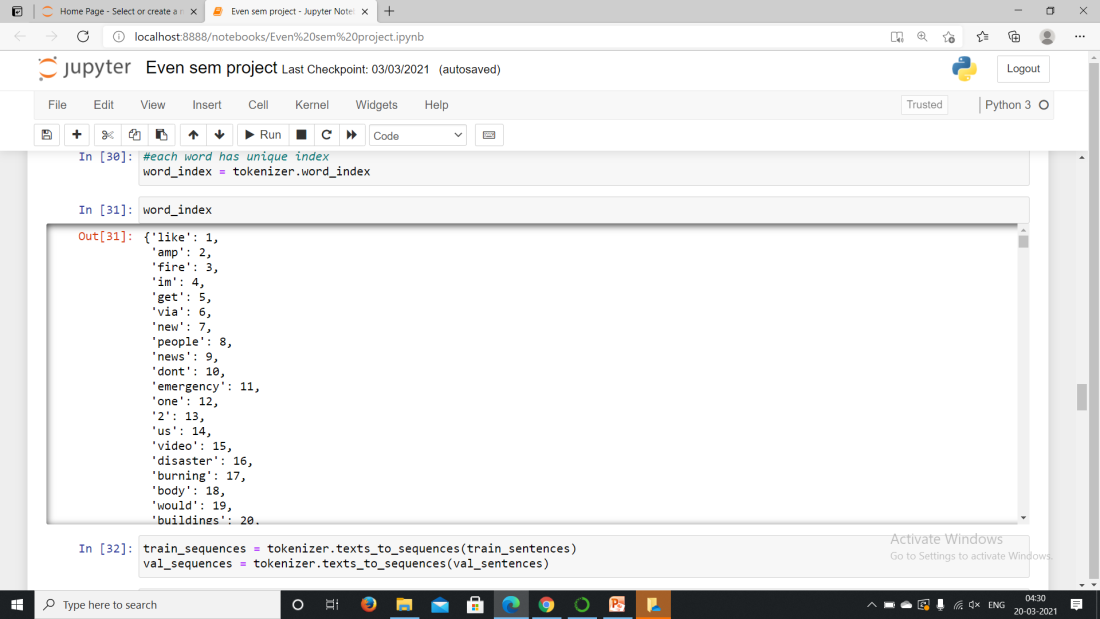


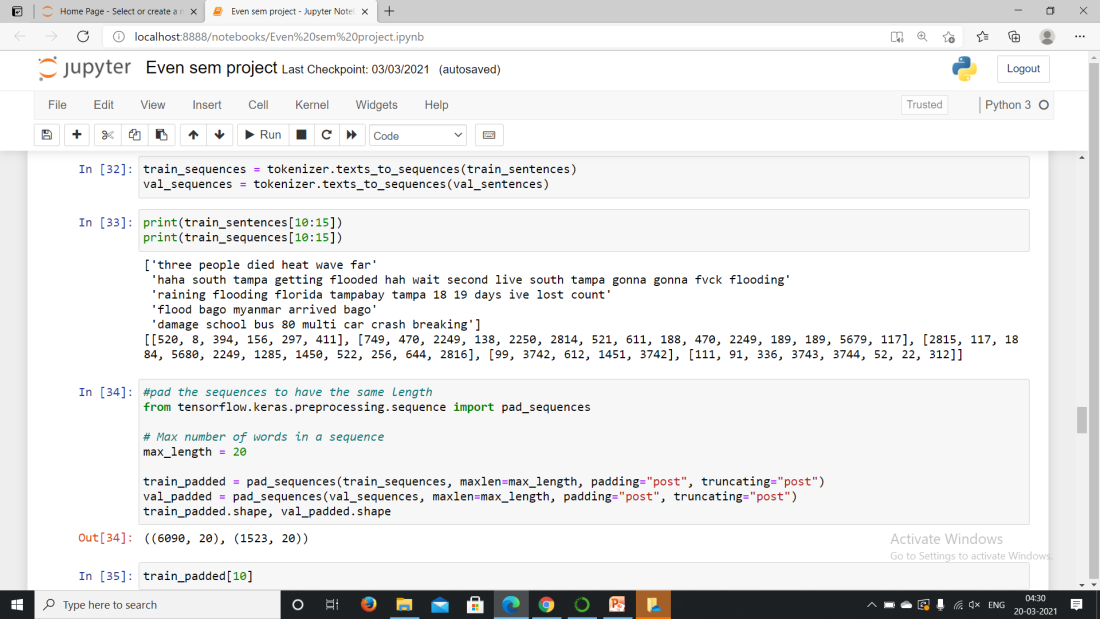


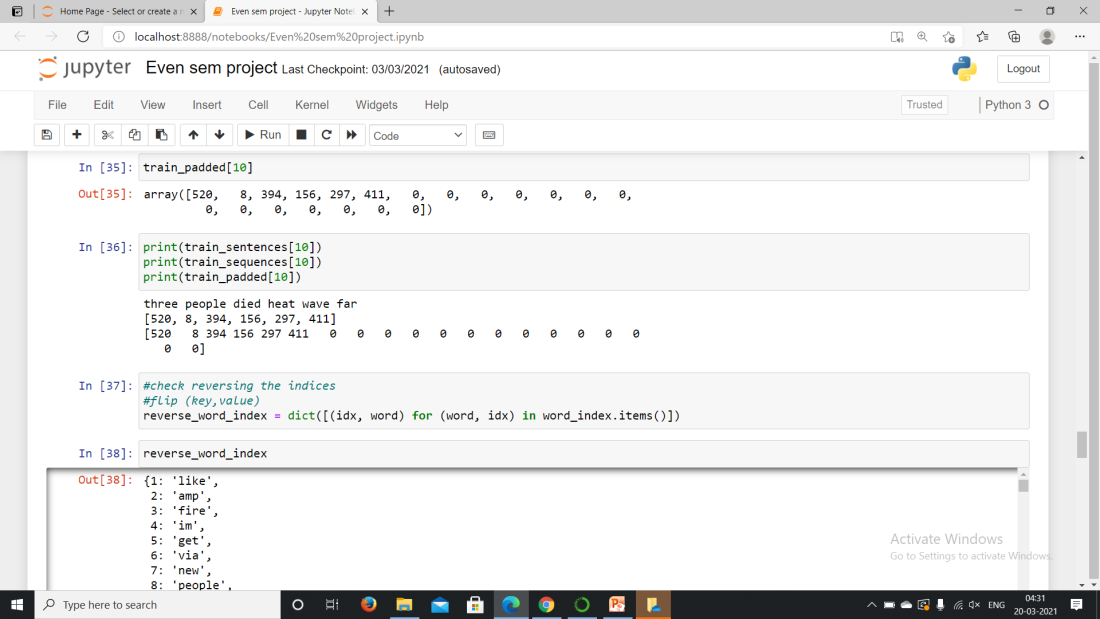


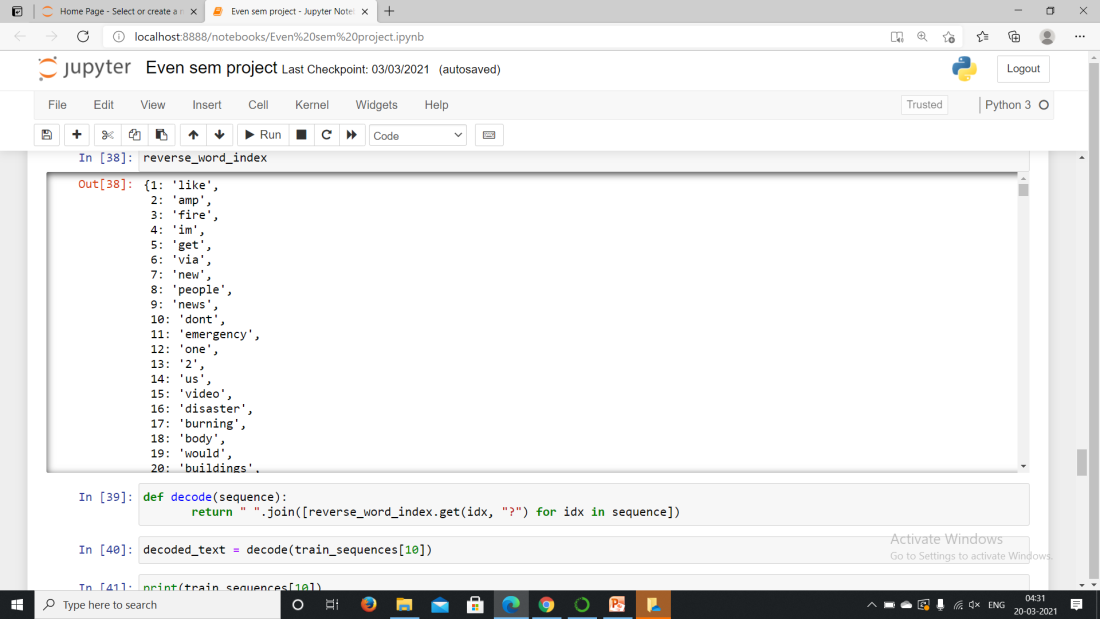


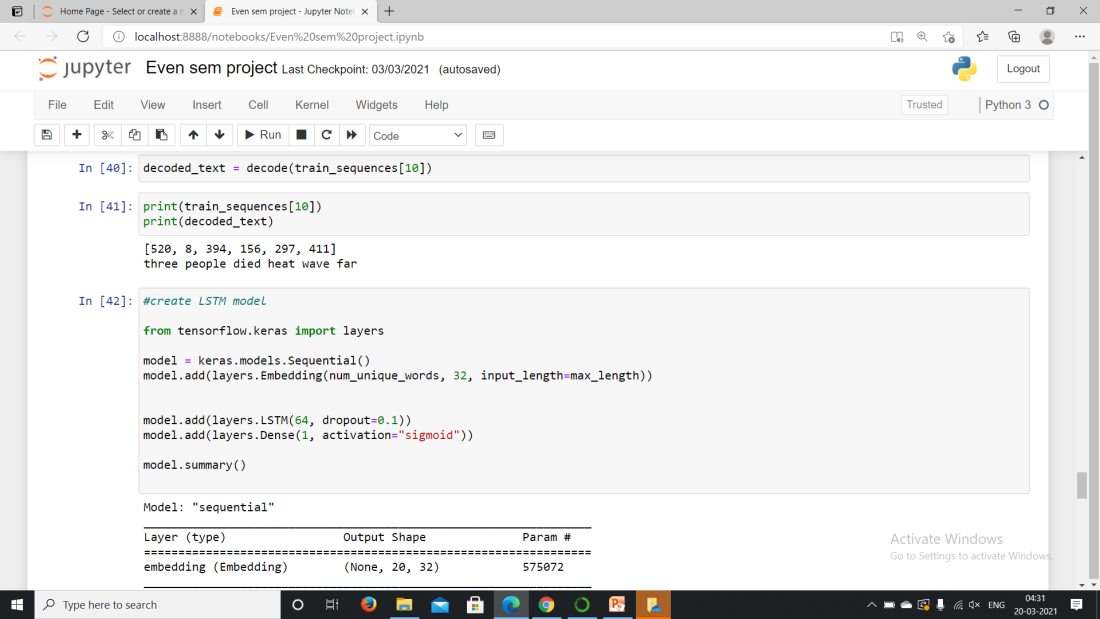


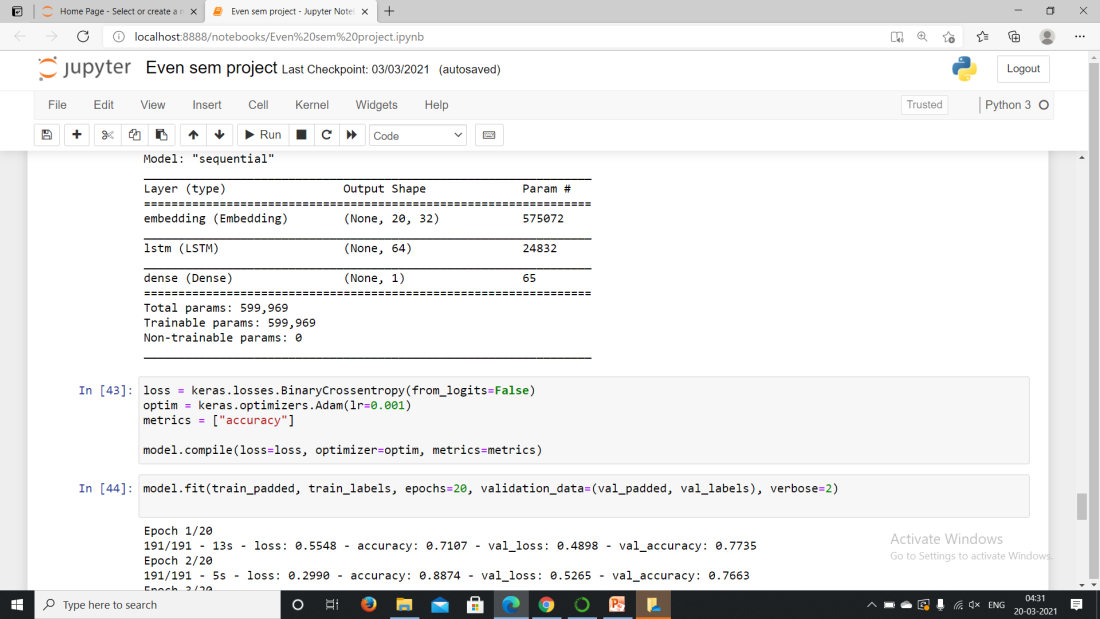


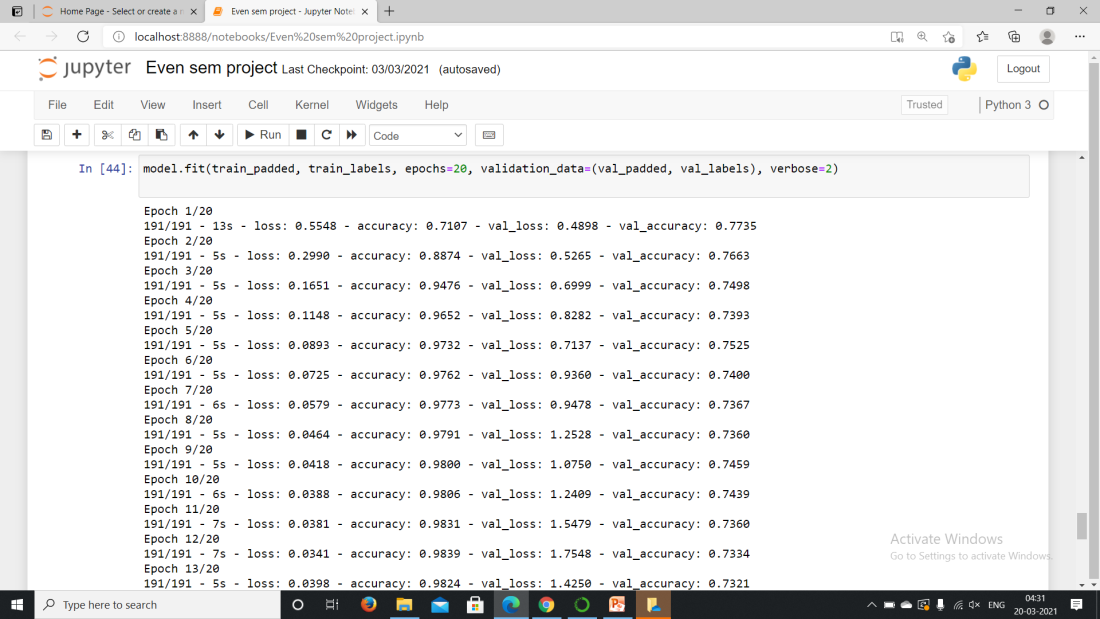


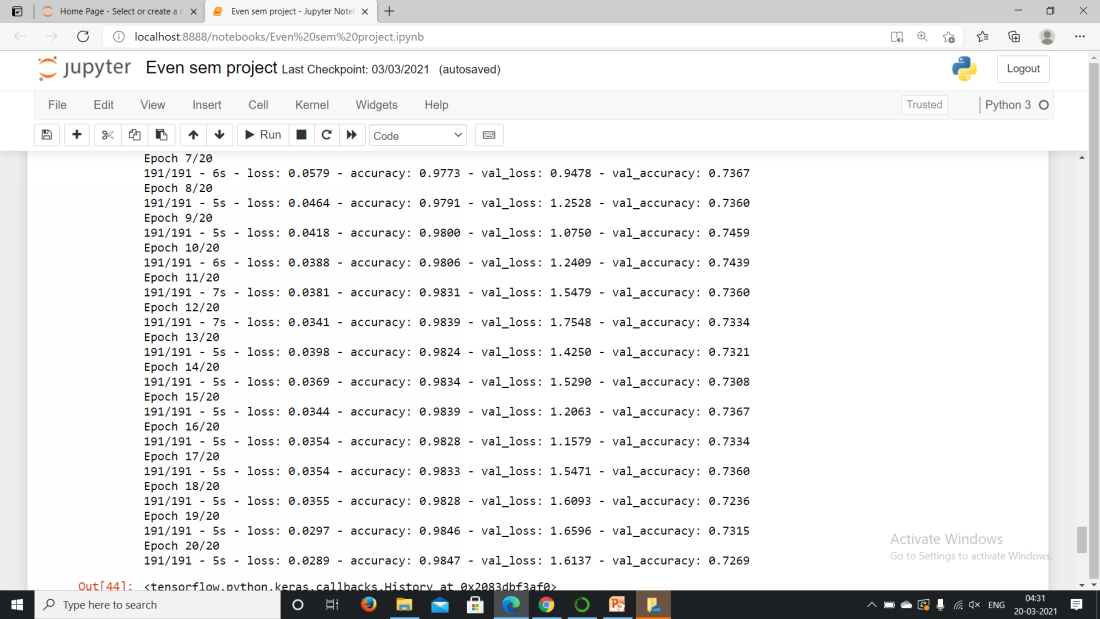












**6: DISCUSSION OF RESULTS**

In this project the aim is to predict whether the news is real or fake by taking datasets from kaggle. To perform this project we used python as working environment. We used some of the data preprocessing to remove the missing values and noise in the data because LSTM itself had inbuilt preprocessing functions. After preprocessing we split the data into train and validation datasets. By using Long short term memory classification algorithm prediction is done. The accuracy is about 98.47%.

**7. CONCLUSION:**

We applied the LSTM (Long short-term memory) algorithm and predicted the accuracy by including limited input layers. We observed in our implementation that LSTM had the best preprocessing techniques which had inbuilt functions and dropout-processing using its layers. The accuracy for the algorithm is about 98.47%. So, we conclude that LSTM is best for predicting fake news using the data mining technique.

**8: REFERENCES**

1. Diana Elizabeth Roy, Prannay S Reddy , "A Study on Fake News Detection Using Naïve Bayes, SVM (Support Vector Machine), Neural Networks and LSTM.", Jour of Adv Research in Dynamical & Control Systems, Vol. 11, 06-Special Issue, 2019.

2. Pallavi B. Petkar, S. S. Sonawane , "Fake News Detection: A Survey of Techniques ", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-9, July 2020.

3. Monika Verma, Sanjeev Sofat ,"Techniques to Detect Spammers in Twitter- A Survey ", International Journal of Computer Applications 85(10):27-32, January 2014.

4. Kelly Stahl ,"Fake news detection in social media " B.S. Candidate, Department of Mathematics and Department of Computer Sciences, California State University Stanislaus, 1 University Circle, Turlock, CA 95382 Received 20 April, 2018; accepted 15 May 2018.

5. Costel-Sergiu Atodiresei\*, Alexandru Tănăselea, Adrian Iftene , "Identifying Fake News and Fake Users on Twitter " Available online 28 August 2018. Published by : Elsevier Ltd.

6. Shlok Gilda , "Evaluating Machine Learning Algorithms for Fake News Detection" 2017 IEEE 15th Student Conference on Research and Development (SCOReD), December 2017, pp.110-115.

7. Pradheepan Raghavan , "Fraud Detection using Machine Learning and Deep Learning", 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)

8. Adi Saputra, Suharjito, "Fraud Detection using Machine Learning in e-Commerce", Published in International Journal of Advanced Computer Science and Applications(IJACSA), Volume 10 Issue 9, 2019.

9. Onam Bharti, Mrs. Monika Malhotra, ”Sentiment Analysis on Twitter Data”, International Journal of Advanced Research in Computer Engi- neering Mobile Computing, IJCSMC, Vol.5, Issue.6,June 2016, pg.601- 609.

10. Abdullah Alsaeedi , Mohammadd Zubair Khan ,"A study on sentiment Analysis Techniques of Twitter Data" Published in International Journal of Advanced Computer Science and Applications(IJACSA), Volume 10 Issue 2, 2019.