**p**

**Fake news detection using nlp**

Phase 5:project documentation and submission

Team member :

* L. Kowsalya
* S. Jeshiya
* V. Sharmila
* D. Aanandhavalli

Topic :Document the project and prepare it’s for submission .

Abstract :

* The spreading of fake news has given rise to many problems in society. It is due to its ability to cause a lot of social and national damage with destructive impacts. Sometimes it gets very difficult to know if the news is genuine or fake.
* Therefore it is very important to detect if the news is fake or not. \"Fake News\" is a term used to represent fabricated news or propaganda comprising misinformation communicated through traditional media channels like print, and television as well as non-traditional media channels like social media.
* Techniques of NLP and Machine learning can be used to create models which can help to detect fake news. In this paper we have presented six LSTM models using the techniques of NLP . The datasets in comma-separated values format, pertaining to political domain were used in the project.
* The different attributes like the title and text of the news headline/article were used to perform the fake news detection. The results showed that the proposed solution performs well in terms of providing an output with good accuracy, precision and recall.
* The performance analysis made between all the models showed that the models which have used GloVe and Word2vec method work better than the models using TF-IDF. Further, a larger dataset for better output and also other factors such as the author ,publisher of the news can be used to determine the credibility of the news.
* Also, further research can also be done on images, videos, images containing text which can help in improving the models in future.

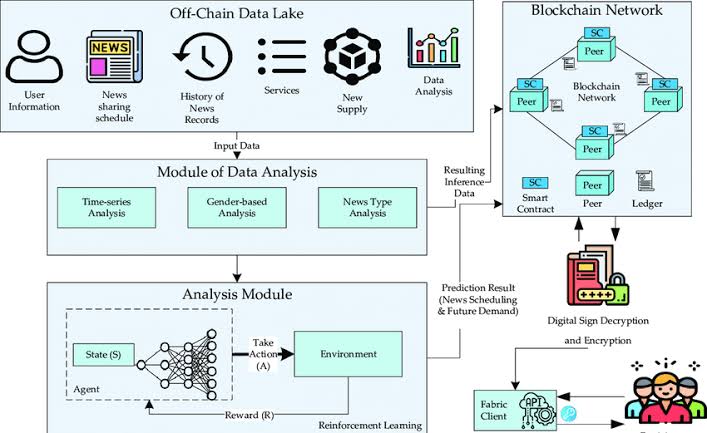
Introduction :

* Fake news is false or misleading information presented as news. The proposed study uses machine learning and natural language processing approaches to identify false news—specifically, false news items that come from unreliable sources.
* Fake news and disinformation are ongoing problems that may be found all around us in biased software that amplifies just our viewpoints for a "better" and smoother user experience.



* Fake news and misinformation are becoming more of a problem as the internet and social media platforms become more main stream. A common goal of fake news is to harm someone or something's reputation or to profit through advertising. The propagation of these ideas may have been influenced by a variety of factors, but they all present humanity with the same underlying problem:
* a misunderstanding of what is real and what is false. This confusion could result in additional problems, such a medical emergency.

Block chain of fake news detection using nlp :



Design and thinking :

This project involves 6 steps. They are Data source, Data preprocessing, Feature Extraction, Model selection, Model Training, Evaluation.

1. Data source

* Gather a diverse dataset of new article or social media posts categorizing them as either real or fake .
* You’ll need labeled data to train and evaluate the NLP model

But in our case we are already provided by datasets for KAGGLE platform .

* Dataset link:

<https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset>

The dataset contain a list of articles considered as “ fake “news.

* They contain 4 columns and 17903 unique values. The dataset in the columns based on

1.Tittle

2.text

3.subject

1. Data pre-processing

* Clean and pre-process the text data by removing stop words, punctuation and special characters

In this part include 3 things

1.Text cleaning

✓ Remove HTML tags ,if any from the text.

✓ convert text to lowercase to ensure uniformity

✓ Handle or remove any non standard characters,puncutation numbers.

2.Tokenization

✓Split the text into individual words tokens.

✓This allows the model to processes the text at the word level.

3.Stop word removal

✓ Remove common stop words (eg., “the”,”and”,”in”)that do not carry’s

Significant

1. : Feature Extraction

• Transform the convert text into numerical features that the common model machine learning model.

• It methods include TF-IDF(Term frequency -inverse Document frequency)

1.Tf-IDF : it measures the importance of a word within document relatives to its importance in the entire corpus.

• .It balances the frequency of a word in a document with its rarity in the corpus making suitable to its identifying signifying terms

2. Word embedding : word. Embedding like word2vec ,glove and fast test capturing

Semantic relationship between word

• These embedding provide dense vector representation of vector representation of words and you can average or concatenate them to represent documents .

1. Model selection

• In this part we are choosing a suitable model for Random forest for fake news detection.

• Model selection involves 4 steps

✓. It is a group of decision tree from a subset of randomly selected a training dataset .

✓. It combines a multiple tree to form a forest hence the random forest to give high accuracy

To the proposed the model

✓.This type of random forest is ensemble learning.

1. .Model training

• Here, we first create an instance of the random forest model ,with the hyper parameter .

• We then fit this to our training data we pass both the targets variables ,so the model can learn

• At this point we have a trained Random forest model ,but we need to find out whether it is making fake news prediction

7 Steps involved in this model training

✓. Import the Skirt-learn python libraries

✓.import the data

✓.Read/clean/adjust the data

✓.Train the classifier model of the training data

✓.create the Random forest model object

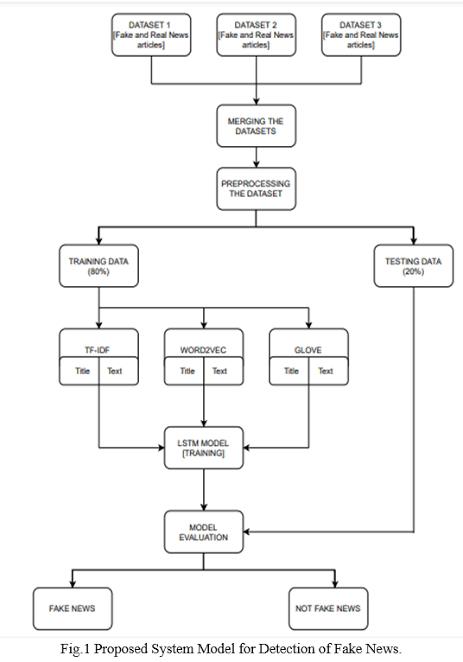
✓. Fit the model

✓.predict

6.Evaluation

* After a training a data evaluate the final random forest model and on the dataset to obtained an unbiased estimate to its performance
* Use appropriate evaluation such as acuracy,precision, recall ,f1-score, and Roc-AUC

Process of fake news detection using nlp mind map



Library using fake news detection using nlp

There are several libraries in Python commonly used for fake news detection using Natural Language Processing (NLP).

Some popular ones include.

NLTK (Natural Language Toolkit): Useful for basic NLP tasks such as tokenization, stemming, tagging, parsing, and more.

Scikit-learn: Provides efficient tools for data mining and data analysis, including various algorithms for classification, regression, clustering, and dimensionality reduction.

Tensor Flow and Keras: Powerful for building and training deep learning models, including those for text classification and sentiment analysis.

PyTorch: Another popular deep learning library that can be used for tasks such as text classification and sentiment analysis.

Genism:

Ideal for topic modeling, document indexing, and similarity retrieval with large corpora.

Transformers (Hugging Face): Particularly useful for implementing transformer-based models like BERT, GPT-3, and others for tasks such as text classification and language understanding.

Innovative technique using during the phase development :

innovative technique for fake news detection using NLP involves leveraging deep learning models, such as transformers, along with various techniques like:

Pretrained Language Models:

* Utilizing pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pre-trained Transformer) for understanding the context and semantics of news articles, thereby capturing subtle linguistic nuances indicative of fake news.
* Multimodal Analysis: Integrating multiple modalities, including text, images, and metadata, to identify inconsistencies between textual content and accompanying multimedia elements, providing a more comprehensive analysis for fake news detection.

Explainable AI Techniques:

* Incorporating explainable AI techniques to interpret and highlight the key features that contribute to the model's prediction, allowing users to understand why a particular piece of content has been classified as fake or genuine

Contextual Embeddings:

* Employing contextual embeddings like ELMO (Embeddings from Language Models) or RoBERTa (A Robustly Optimized BERT Pretraining Approach) to capture complex linguistic patterns and dependencies, enabling the model to better discern between authentic and fabricated information.

Cross-lingual Analysis:

* Extending the analysis to multiple languages by leveraging cross-lingual

### [Things you need to install](https://github.com/mohammed97ashraf/Fake_news_Detection#things-you-need-to-install)

1. Python 3.9
   * This setup requires that your machine has python 3.9 installed on it. you can refer to this url <https://www.python.org/downloads/> to download python. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: <https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/>.
2. You will also need to download and install the required packages after you install python
   * Sklearn (scikit-learn)
   * numpy
   * Pandas
   * matplotlib
   * seaborn
   * NLTK
   * Joblib
   * flask
3. To install the Packages

pip3 install -U scikit-learn

pip3 install numpy

pip3 install Pandas

pip3 install matplotlib

pip3 install seaborn

pip3 install nltk

pip3 install flask

pip3 install joblib

Given dataset:

In1

import numpy as np *# linear algebra*

import pandas as pd *# data processing, CSV file I/O (e.g. pd.read\_csv)*

import plotly.express as px

import plotly.graph\_objs as go

from plotly.subplots import make\_subplots

import nltk

from nltk.corpus import stopwords

import tensorflow as tf

from tensorflow.keras.optimizers import Adam

from tensorflow.keras.callbacks import ModelCheckpoint

from sklearn.model\_selection import train\_test\_split

from transformers import AutoTokenizer, TFAutoModelForSequenceClassification

Out1:

True

In2

fake\_news\_path = "/kaggle/input/fake-and-real-news-dataset/Fake.csv"

real\_news\_path = "/kaggle/input/fake-and-real-news-dataset/True.csv"

In3

fake\_df.isnull().sum()

Out2:

title 0

text 0

subject 0

date 0

dtype: int64

real\_df.isnull().sum()

Out3:

title 0

text 0

subject 0

date 0

dtype: int64

Preprocessing the dataset:

* Tokenization:-

Tokenization is the process of breaking down a piece of code into smaller units called tokens

import nltk

from nltk.tokenize import RegexpTokenizer

# Create tokens out of alphanumeric characters

tokenizer = RegexpTokenizer(r'\w+')

tokens = tokenizer.tokenize("I think pineapple pizza is gross and not worth $15!")

print(tokens)

# ['I', 'think', 'pineapple', 'pizza', 'is', 'gross', 'and', 'not', 'worth', '15']

* Lemmatization:

:import nltk

from nltk.stem import 8

from nltk.corpus import wordnet

lemmatizer = WordNetLemmatizer()

# Convert the nltk pos tags to tags that wordnet can recognize

def nltkToWordnet(nltk\_tag):

if nltk\_tag.startswith('J'):

return wordnet.ADJ

elif nltk\_tag.startswith('V'):

return wordnet.VERB

elif nltk\_tag.startswith('N'):

return wordnet.NOUN

elif nltk\_tag.startswith('R'):

return wordnet.ADV

else:

return None

# Lemmatize a list of words/tokens

def lemmatize(tokens):

pos\_tags = nltk.pos\_tag(tokens)

res\_words = []

for word, tag in pos\_tags:

tag = nltkToWordnet(tag)

if tag is None:

res\_words.append(word)

else:

res\_words.append(lemmatizer.lemmatize(word, tag))

return res\_words

* Stemming

# Using Porter Stemmer implementation in nltk

from nltk.stem import PorterStemmer

stemmer = PorterStemmer()

def stem(tokens):

return [stemmer.stem(token) for token i tokens]

tokens = ['jumped', 'jumps', 'jumped']

print(stem(tokens))

# ['jump', 'jump', 'jump']

Fake news detection using nlp benefits:

1. **Enhanced Information Credibility:** By accurately identifying and flagging fake news, NLP helps to improve the credibility and reliability of the information circulating in the media and online platforms, thus fostering a more informed and discerning society.
2. **Mitigation of Misinformation:** NLP-based fake news detection systems contribute to the mitigation of misinformation by promptly identifying and addressing false or misleading content, thereby reducing the negative impact of false information on public opinion and decision-making.

Conclusions :

* In conclusion, fake news detection using NLP is a vital tool for combating the proliferation of misinformation and fostering a more informed and discerning society
* By leveraging advanced NLP techniques and models, such as pretrained language models, multimodal analysis, and semantic similarity analysis, we can effectively identify and flag false or misleading information, thus safeguarding the credibility and reliability of the information circulating in the media and online platforms.
* As we continue to refine and innovate within the field of NLP, it is essential to prioritize the development and implementation of robust and adaptable systems that can effectively discern between authentic and fabricated information,
* thus contributing to the creation of a more resilient, responsible, and well-informed global community