Final Project

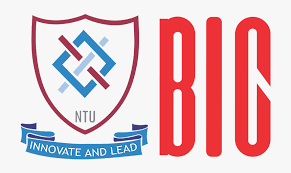
BS SOFTWARE ENGINEERING

6th SEMESTER

**NOOR ul HUDA 20-NTU-CS-1133**

**AFFAN TAHIR 20-NTU-CS-1099**

**ALEESHA RUSTAM 20-NTU-CS-1094**



**SUBMITTED TO : Mam Sajida Parveen**

**DATE : 12-06-2023**

**FAKE NEWS PREDICTION**

**INTRODUCTION:**

Are you looking for ways to grow the user base for your mobile application? Then you have arrived at the right place. Here you will find a curated collection of landing page HTML templates that will help you build an engaging online presentation for your mobile app and convince visitors to become loyal paying users. The proliferation of fake news has become a pressing issue in the digital age, posing significant challenges to individuals, society, and the media. Fake news can spread rapidly and have detrimental effects, including influencing public opinion, exacerbating social divisions, and undermining trust in reliable information sources. In response to this critical problem, the development of effective fake news detection systems has gained immense importance.

The objective of this project is to develop an unsupervised fake news detector that can automatically identify and distinguish between real and fake news articles. Unlike supervised approaches that require labeled training data, the unsupervised approach leverages techniques such as natural language processing, dimensionality reduction, and unsupervised learning algorithms to detect patterns and anomalies in the dataset.

**SIGNIFICANCE OF THE PROJECT**

The significance of this project lies in its potential to provide a valuable tool for media organizations, fact-checkers, and online platforms to combat the spread of misinformation. The unsupervised approach offers advantages such as scalability and adaptability to evolving forms of fake news, making it a promising avenue for addressing this challenging problem.

Through this project, we aim to shed light on the capabilities and limitations of unsupervised fake news detection, analyze the performance of the developed system, and provide insights into the effectiveness of such approaches in tackling the issue of fake news.

* **Problem Statement**: The problem is to develop an unsupervised fake news detection system that can automatically identify deceptive news articles without the need for labeled training data.
* **Motivation:** The motivation behind this project is to address the growing issue of fake news by leveraging unsupervised machine learning techniques to detect and flag potential instances of deception in news articles, aiding in promoting information integrity and combating the spread of misinformation.
* **Dataset:** <https://github.com/DanishJameel/Fake_News_Detection_Web_Application>

**LIBRARIES**

* **PANDAS**

Powerful data manipulation and analysis library. It provides data structures and functions for efficiently handling structured data.

* **MATPLOTLIB**

Matplotlib is a plotting library that provides a comprehensive set of tools for creating static, animated, and interactive visualizations in Python.

* **DJANGO**

Django is a high-level web framework for building web applications in Python. It provides a robust set of tools, templates, and features for handling web development tasks.

* **NLTK**

NLTK is a library for natural language processing (NLP).NLTK is commonly used for text processing, text mining, and building NLP applications.

* **SCIKIT-LEARN**

Scikit-learn is a popular machine learning library in Python. It provides a wide range of algorithms and tools.

* **PICKLE**

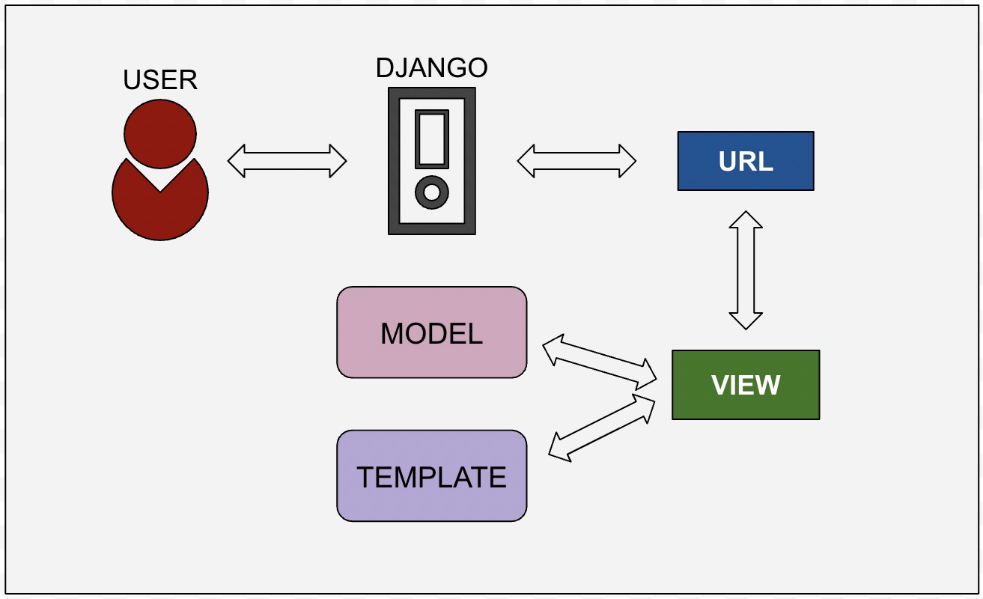
Pickle is a standard library in Python that allows you to serialize Python objects into a binary format and save them to a file.

* **SEA BORN**

Seaborn is a data visualization library built on top of Matplotlib. It provides a higher-level interface for creating aesthetically pleasing statistical graphics.

**ARCHITECTURE**

The **Model-View-Template** (MVT) architecture is a web development pattern that separates different components of a web application. The Model represents the data and business logic, the View handles the presentation logic, and the Template defines the structure of the user interface.

****

**METHODOLOGY**

1. **Word Tokenization**

Converting textual data into numerical vectors to enable machine learning algorithms to process them, commonly using techniques like TF-IDF.

1. **Gussian Mixture Models**

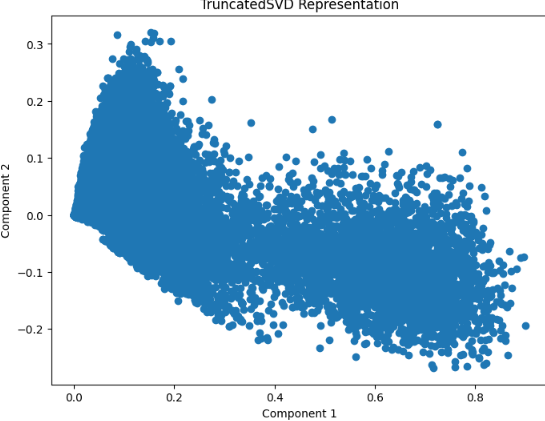
Probabilistic models that represent a dataset as a mixture of Gaussian distributions, used to model the underlying distribution of real and fake news articles.

1. **Anamoly Detection**

Identifying data points that deviate significantly from the norm or expected behavior, applied to detect potential outliers or suspicious instances in the context of fake news detection.

**ALGORITHMS**

**Truncated Singular Value Decomposition (SVD)** is a dimensionality reduction technique that plays a crucial role in the fake news detection system. In the context of this project, Truncated SVD is applied to the TF-IDF vectors obtained from the vectorization process. By retaining only the most important singular values and corresponding singular vectors, Truncated SVD reduces the dimensionality of the feature space. This reduction helps in eliminating noise and less informative features, while retaining the main patterns and structure of the data. Through this process, Truncated SVD aids in improving computational efficiency and enhancing the performance of subsequent algorithms. The number of components or clusters you want to model are fixed. In this case, we have set it to 2, indicating that we want to model real news and fake news as separate components.

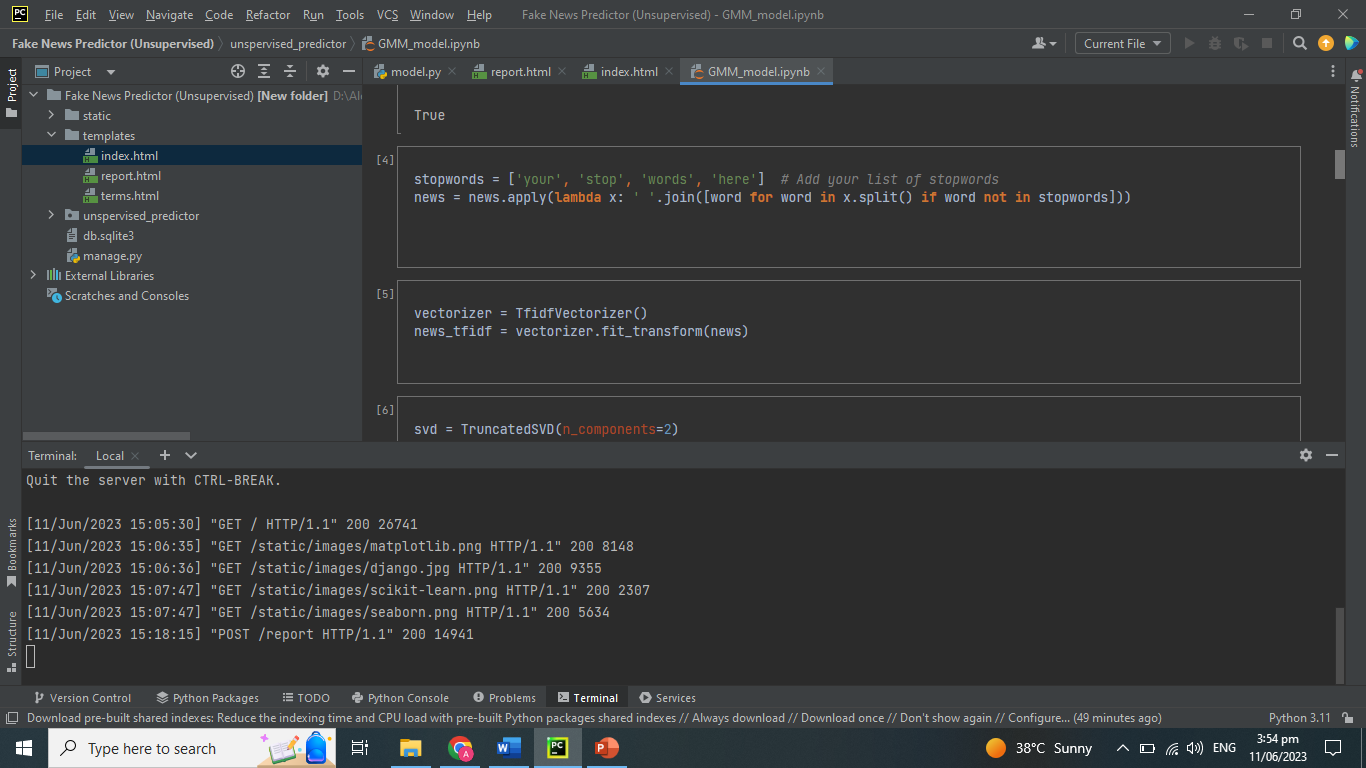


**Isolation Forest** is an unsupervised machine learning algorithm used for outlier detection. It was introduced in 2008 as a way to efficiently identify anomalies in data. The main idea behind Isolation Forest is that anomalies are typically few and different from the majority of the data, making them easier to isolate. The algorithm constructs a collection of binary trees, called isolation trees, to separate anomalies from normal data points.

**PREPROCESSING AND VECTORIZATION**

In preprocessing we clean our dataset. First of all we remove stop words from the news and get the main content that we needed and then do vectorization.

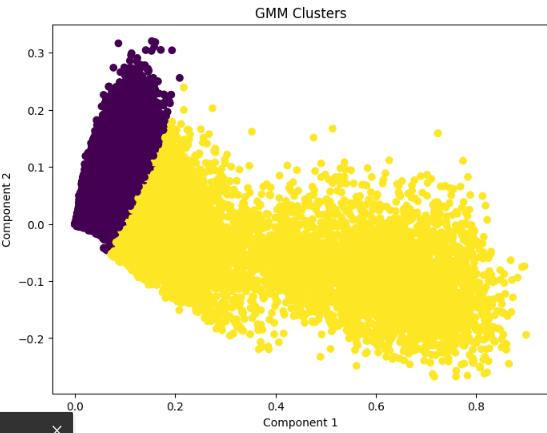
**TfidfVectorizer** from the sklearn.feature\_extraction.text module is used to convert the text data into numerical vectors. TF-IDF stands for Term Frequency-Inverse Document Frequency, and it represents the importance of a word in a document relative to the entire dataset.



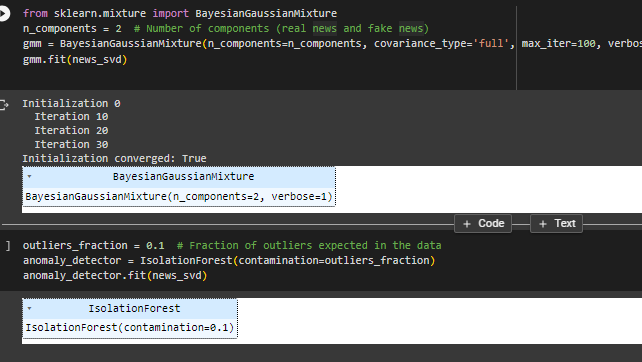
**GAUSSIAN MIXTURE MODELS**

GMM are probabilistic models that represent a dataset as a mixture of Gaussian distributions. In the context of fake news detection, GMM is used to model the underlying distribution of real news and fake news articles based on their features.

By estimating the parameters of each Gaussian component, such as the mean and covariance, GMM learns the distribution of the data and assigns probabilities to each data point for belonging to each component. For clustering, the GMM assigns each data point to the Gaussian component with the highest responsibility, effectively grouping similar data points together.

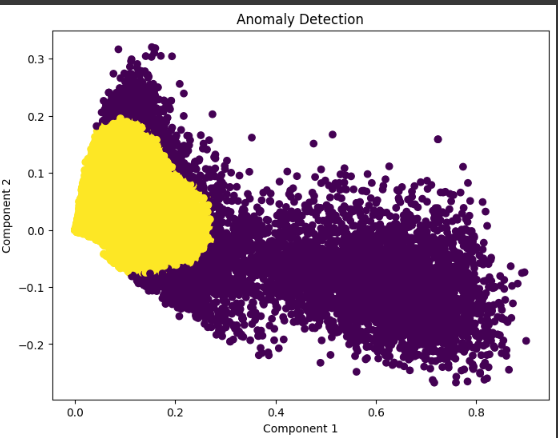


**CODE SNIPPETS IMPLEMENTED MODELS**



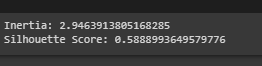
**ANOMALY DETECTION**

* Focuses on identifying unusual or anomalous patterns in data.
* Typically operate in an unsupervised manner.
* It includes statistical methods, clustering-based methods, and density estimation techniques.

****

**EVALUATION TECHNIQUES**

To evaluate the performance of the clustering and anomaly detection processes, two evaluation metrics are commonly used: Inertia and Silhouette Score. Inertia measures the within-cluster sum of squared distances and provides insight into the compactness or coherence of the clusters. Lower inertia values indicate better clustering performance, suggesting that the clusters are more tightly packed. On the other hand, the Silhouette Score assesses the quality of clustering by calculating the average silhouette coefficient for each sample. The silhouette coefficient measures how similar a sample is to its own cluster compared to other clusters. A higher Silhouette Score indicates well-separated clusters, signifying a better clustering result. These evaluation metrics help in assessing the effectiveness of the unsupervised fake news detection system and provide quantitative measures of its performance.



**SILHOUETTE SCORE**

The silhouette score measures how close each sample in one cluster is to the samples in the neighboring clusters. It ranges from -1 to 1, where a higher value indicates better-defined clusters. A silhouette score close to 1 indicates that the samples are well-clustered, with clear separation between clusters.

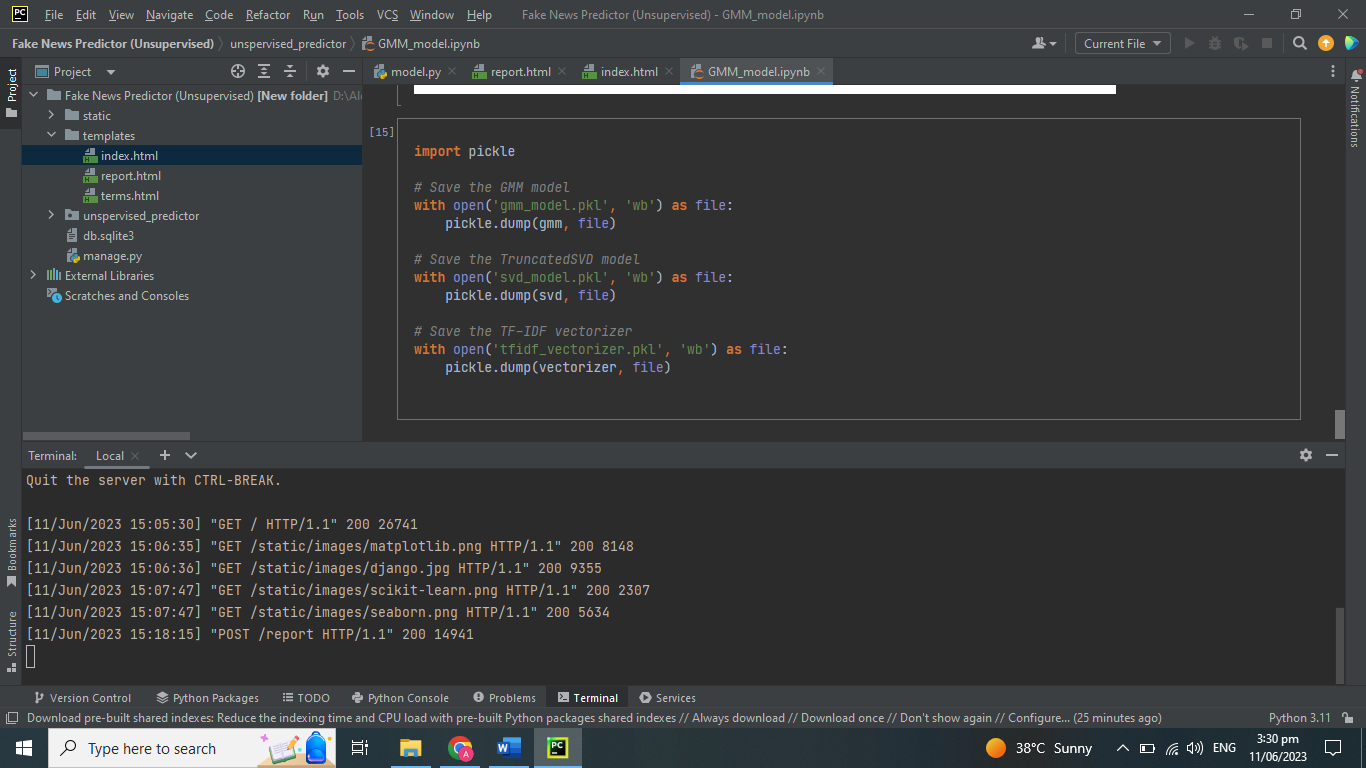
**INERTIA**

Inertia represents the sum of squared distances of samples to their closest cluster center. It measures how compact and well-separated the clusters are. A lower inertia value indicates that the data points within each cluster are closer to their respective cluster centers, suggesting better-defined and more cohesive clusters.

**SAVING DATA**

The code snippet you provided demonstrates how to save the GMM model, TruncatedSVD model, and TF-IDF vectorizer using the pickle module in Python. Pickle is a standard library in Python that allows you to serialize Python objects into a binary format and save them to a file.

In the code, you open each file using the open() function with the file name and the mode 'wb' (write binary) to indicate that you want to write binary data. Then, you use the pickle.dump() function to serialize and save the respective objects to the files.

****

**PROGRAMMING INTERFACE PERCENTAGES**

* Django: 40
* Jupyter NoteBook: 20
* Web: 40

**\*\*\* FOR RUNNING THE APPLICATION PERFORM BELOW STEPS BEFORE \*\*\***

1. create a directory with any name like FakeNewsPredictor
2. open folder in VS code and then open terminal for the directory or simply open Command prompt in the folder
3. create a virtual Enviroment using following command:

* python.exe -m pip install --upgrade pip       **#Upgrade your pip**
* python -m venv myenv                      **#create a virtual enviroment**
* myenv/scripts/activate                    **# to activate venv in cmd or terminal**
* source myenv/scripts/activate             **# to avtivate myenv in gitbash**

1. then install DJANGO and below libraries:

* pip install Django
* pip install Pandas
* pip install scikit-learn

1. after that change the directry using command

* cd fake\_news\_detector

1. run the project now

* python manage.py runserver