Fake News Detection

(DigiAuth)

By

Zuha Tassawar

Hurrairah Imran

Mughees Anwar Lone

Umair Ahmed

Supervised by:

Dr. Saddaf Rubab

Submitted to the faculty of Department of Computer Software Engineering,
Military College of Signals, National University of Sciences and Technology, Islamabad,
in partial fulfillment for the requirements of B.E Degree in Software Engineering.

June 2022

In the name of Allah, the Most Benevolent, the Most Courteous

CERTIFICATE OF CORRECTNESS AND APPROVAL

This is to officially state that the thesis work contained in this report

"Fake News Detection"

is carried out by

Zuha Tassawar, Hurrairah Imran, Mughees Anwar, Umair Ahmed

under my supervision and that in my judgement, it is fully ample, in scope and excellence, for the degree of Bachelor of Software Engineering in Military College of Signals, National University of Sciences and Technology (NUST), Islamabad.

Approved by

Supervisor
Dr. Saddaf Rubab

Date: _____

DECLARATION OF ORIGINALITY

We hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification in either this institute or anywhere else.

ACKNOWLEDGEMENTS

Allah Subhan'Wa'Tala is the sole guidance in all domains.

Our parents, colleagues and most of all supervisor, Dr. Saddaf Rubab, without your guidance.

The group members, who through all adversities worked steadfastly.

Plagiarism Certificate (Turnitin Report)

This thesis has 7% similarity index. Turnitin report endorsed by Supervisor is attached.

Zuha Tassawar

00000248736

Hurrairah Imran

00000278766

Mughees Anwar Lone

00000280710

Umair Ahmed

00000240958

Signature of Supervisor

FYP Report

ORIGINALITY REPORT

7%	7%	1%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	vanrijmenam.nl Internet Source	3%
2	deepai.org Internet Source	1%
3	Submitted to Perdana University Student Paper	<1%
4	Submitted to Higher Education Commission Pakistan Student Paper	<1%
5	www.essaysauce.com Internet Source	<1%
6	hawk.csc.vill.edu Internet Source	<1%
7	www.scale.szamalk.hu Internet Source	<1%
8	Arpanjeet Sandhu, Sabah Mohammed. "Detecting Anomalies in Logs by Combining NLP features with Embedding or TF-IDF",	<1%

	Institute of Electrical and Electronics Engineers (IEEE), 2022 Publication	
9	id.scribd.com Internet Source	<1%
10	Submitted to National College of Ireland Student Paper	<1%
11	Jatin Gupta. "The Accuracy of Supervised Machine Learning Algorithms in Predicting Cardiovascular Disease", 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST), 2021 Publication	<1%
12	Submitted to University of Leicester Student Paper	<1%
13	kbtcoe.org Internet Source	<1%
14	www.slideshare.net Internet Source	<1%
15	"Business Information Systems Workshops", Springer Science and Business Media LLC, 2019 Publication	<1%
16	businessdocbox.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Abstract

Social network developers and researchers are facing a big challenge in the form of rumors and misleading information. The verification of the reliability and credibility of is a vital requirement for social media consumers. Thus, Blockchain is a powerful tool and it's secure and tamperproof p2p design and immutability provides this magical solution.

The motive of this project is to diminish the spread of misinformation and be a means for the true information to be visible to the public. This also promotes the true essence of journalism and collaborates with the journalists and other platforms to combat fake news on the web.

The main idea behind the project is to create a system to flag news as fake or real. The system will take text data by scrapping the online news sites(predefined) and after pre-processing the data, NLP techniques would be applied on it to classify it as fake or true. The next module will be about registering the authenticators into the blockchain ledger. The ML classified news will be inputted into the ledger where the authenticators will authenticate it and flag it as fake or true. This flaggednews will be visible to the public through the web screen along with the reputation score of the authenticator provided by a consensus algorithm.

Table of Contents

List of Figures

Chapter 1: Introduction

A lot of attention is being provided to this rise of "fake" news in recent years; covering everything from misleading content to satire to those articles that have no ounce of truth in them. In 2018, 28 countries were surveyed in the Edelman Trust Barometer. The reports mention that

- 70% are concerned with fake news as "weapon"
- 63% cannot differentiate journalistic news from falsehoods
- 59% were not sure if the information they see in the media is reliable

One of the biggest concerns are the "deepfakes" - tampered audios and videos that intend in mimicking celebrities or politicians. For example, Buzzfeed in 2018, with voiceover of Jordan Peele, created a deepfake of Barak Obama, just to raise awareness of this manipulative technology.

Language generation models such as GPT-3, have these amazingly advanced qualities that include generating paragraphs of “realistic” text in English. It can even answer human questions correctly up to some level. But these qualities and capabilities can help malicious agents to use it for evil. Creating articles of false news with realistic writing is just an example.

Although GPT-3 and deepfakes are only a problem in theory so far, with these rapid advancements fake news is becoming a bigger issue than it is perceived. But with the entrance of blockchain: a peer-to-peer technology using distributed ledgers to record every single bit of information and making it completely transparent and virtually impossible tamper or falsify and even hack.

If blockchain is an answer for all problems ranging from art and NFTs to cryptocurrencies to identifying counterfeit wines—then what about battling fake news too?

1.1 Overview

For finding fake news, it is important to develop a model that distinguishes between "true" and "fake" news. This will have an impact on the social networking sites and microblogging such as Instagram, Twitter, Facebook and on the instant communication apps like Telegram and WhatsApp. In short, all the places where misleading news is gaining weight and spreading from a national to a global level. False news is just a fabrication that is tampered to be seen as genuine report thus gaining attention and financial gain. Among the 4000 languages spoken, each having its own script, syntax, and style, it is difficult to understand because of the language barriers and all the uses of similes, idioms, and metaphors and so on. By using indigenous language analysis which is a set of artificial intelligence that includes methods of text analysis, modeling, and prediction, we develop a project that uses historical data from newspapers to find out what is true and what is not.

1.2 Problem Statement

In Pakistan which is a third world underdeveloped country; the spread of fake news is getting higher with every passing day. Our news channel spread news with no ounce of authenticity daily, and to overcome this problem we need a solution which can cater this cause and deduce this problem, so we do not stay any behind in the 5^{th} generation warfare.

1.3 Proposed Solution

DigiAuth seeks to diminish the spread of misinformation. It empowers the readers into make informed, and confident factual judgments about the content they see online. It works like two-way authentication. First, AI algorithms are applied to the data, after that our data is imputed into a blockchain ledger where further it is processed for its authenticity.

1.4 Working Principle

The project mainly works on the principles of blockchain for the preservation of data amalgamated with machine learning algorithms for detection model. The project is divided into different modulus and every module is inter-woven with the next module. The list of modules is as under:

- Datasets and annotations

- Dataset training and processing
- Output extraction
- Decision based upon output
- Integration of ML on blockchain using
- Ethereum
- GUI presentation on Web3

1.5 Objectives

1.5.1 General Objectives:

"To build an innovative state of the art software integrated hardware prototype powered by Machine Learning (ML) and Internet Protocol (IP) techniques, providing a smart administrative tool to reduce the traffic congestion problem."

1.5.2 Academic Objectives:

- Development of a smart and intelligent Traffic administration System
- To implement Machine Learning techniques and simulate the results
- To increase productivity by working in a team
- To design a project that contributes to the welfare of society

1.6 Scope

In the field of journalism, whether it be online or offline, fake news is constantly presenting itself as a major issue. That is why many media organizations are constantly on the look for a better solution to tackle this issue. One of the best solutions to this is The Blockchain. This distributed ledger can keep a record of where, when and by whom the news was published, the video or photograph was taken. This provides evidence of all the metadata attached with a single piece of news and along with that allows the users to validate and fact-check information themselves. This not only offers provenance and transparency to the people but also reduces the rate of spread of rumors and misinformation.

1.7 Deliverables

1.7.1 Rating of source

It serves as a rating system to rate the sources by the authenticators according to their credibility of the source. The better the rating of a source the genuine the news.

1.7.2 Object of interest:

It can detect the object of interest by using the same combination of data sets and machine learning techniques. By detecting the object of interest, we mean detecting the fakeness of the news and its rating by authenticators.

1.7.3 Special privileges:

It provides the special privileges to the local audience, which is not common in any system these days, every individual can access to our system on web and check for the authentication of the news.

1.8 Relevant Sustainable Environmental Goals

1.8.1 Decent Work and Economic Growth

Transparency among media leading to fair judgment and selection. Users can make their own judgement about topics and this can allow people to think for themselves.

1.8.2 Industry, Innovation, and Infrastructure

Brands can avoid all the online abuse they are facing and notify their users in case of any misinformation. This can increase trustworthiness between producers and consumers. Social Media outlets can also discourage their users from interacting with fake content, and this can create a real community.

1.8.3 Peace, Justice, and Strong Institutions

The society in which people make judgment by themselves and the spread of misinformation and rumors can be eradicated leads to strong people and peace among the community.

1.9 Structure of Thesis

Chapter 2 contains the literature review and the background and analysis study this thesis is based upon. Chapter 3 contains the design and development of the project.

Chapter 4 introduces detailed evaluation and analysis of the code.

Chapter 5 contains the conclusion of the project.

Chapter 6 highlights the future work needed to be done for the commercialization of this project.

Chapter 2: Literature Overview

A new product is launched by modifying and enhancing the features of previously launched similar products. Literature Review is an important step for the development of an idea to a new product. Likewise, for the development of a product and for its replacement related to reducing the spread of fake news, a detailed study regarding all similar projects is compulsory. Our research is divided into the following points:

- The existing use of blockchain by media and journalist organizations
- Start-ups tackling fake news
- Deep Entity Classification (DEC)
- Research papers

2.1 The existing use of Blockchain in Media Organizations

The problem of fake news exists between two different issues:

1 The misinterpretation of original source or taking it out of context.

2 Media organizations or social sites having a reputation for misinformed reporting.

Media organizations that are in the process of using blockchain are thus in turn addressing the first issue: proving that their media content is authentic. For Example, The News Provenance Project [12], a project of The New York Times that will help fight the spread of fake news.

The project's main aim was to use blockchain to counteract misinformation. For every visual content, context is included with it. For example, a photo taken of one news might be reused to illustrate a completely different news event at a different time and location, so to counteract it, the date, time, name of photographer and other related metadata is added with it to remove the probability of usage of photo for other purposes.

The New York Times, after several consumer interviews came to the following conclusions regarding the best way to combat false and fake news:

- People want facts but they make their own opinion. Questions and prompts will allow user to think for themselves and remove signs of misinformation.
- Details and visual medial and the details about visual media i.e., photographer, date, location, publisher remove the pressure or any outside influence from user and allow them to think for themselves.
- People believe the stories and the skepticism to fact-check some kind of story in the news is just an exception and not a norm.
- Fact-checking should occur before the exposure of story. The longer people are exposed to it the more they believe it.

The New York Times collaborated with IBM to build a proof of concept: a simulated social media website that includes pop-up windows and provides an accurate metadata for every photograph on a user's timeline.

Italy's Agenzia Nazionale Stampa Associata (ANSA) in collaboration with Ernst & Young announced in 2020 to launch a similar initiative "ANSAcheck". It will allow readers to verify the accuracy of any article through a tracking label that provides context like the time and date of when the article was published. They use Ethereum to create smart contracts that link the article to whatever its digital representation is on the blockchain ledger. The blockchain will also keep record of anyone who tampers or modifies the article in it. In this was the publication history is completely transparent to the user.

2.2 The existing use of Blockchain by journalists

While large organization focus on first problem, startups have their focus on the second one that us to combat the spread of fake news. Civil Media, a blockchain startup launched CVL, its own cryptocurrency in 2018 to help incentivize journalistic accuracy.

Instead of focusing on gaining user's attention to generate revenue like other main organizations, they proposed to letting user's vote on the fair sites.

2.3 Start-ups tackling Fake News

The *StartUs Insights Discovery Platform* identified 331 relevant solutions that fight fake news. Figure 1 shows 5 of these that are working in the fierce battle to combat fake news.

Figure 1- Global Startup Heat Map

2.3.1 NewsCheck – Content Review

This gives a scoring platform for the content using the help of both humans and machinery providing fast fact-checking for brands and protecting them for nefarious agents.

2.3.2 Cyabra – Online Brand Abuse

This Israeli startup looks uses NLP and deep learning to look for any misleading information on the social media. They also shield brands from online abuse and thus exposing them to genuine user engagement, and provide services to mitigate real-time online threats to the government entities.

2.3.3 Defudger – Deepfake Detection

This uses Artificial Intelligence and Computer Vision to verify authentic visual content by detecting tampering in them and detects deep fakes. A blockchain model used also helps to verify the authenticity of the content in addition to maintaining records of validated and debunked content.

2.3.4 Primas – Distributed Content

This uses Distributed Trusted Content Protocol (DTCP) to improve the quality of online content by generating metadata of the content. The DTCP tracks the entire network of the data and records that on the blockchain. This helps in content scoring and addressing low quality content from fake news to plagirsm.

2.3.5 Blackbird.AI – Disinformation Countermeasure

This US-based startup uses machine learning and human intelligence to fight misinformation. The AI monitors large volumes of online content and while doing this identifies threat using deep networks. This solution uncovers behavior patterns of the consumers to enable preemptive threat management.

2.4 Deep Entity Classification (DEC)

Fake profiles on social media accounts are one of the most common spread of rumors, and all the tech giants are trying to combat this issue. For example, in 2019 Facebook removed more than a billion accounts that they deemed fake using Machine Learning using a model called Deep Entity Classification (DEC), which detects fake accounts by assessing the patterns in how these

profiles-built connections through Facebook. Reportedly, Facebook has been able to keep fake accounts usage to less than 10% of active users per month.

But how does this work?
Facebook either converts the user-misclassified accounts (accounts for pets, hobbies, etc.) into pages, or uses machine learning to block the Violating accounts (accounts that scam or spam or violate terms of service) before they harm the user i.e., before they are created or become active. But when a fake account has become active, it gets harder to detect it, so Facebook uses Deep Entity Classification (DEC).

It differentiates fake and real users by the deep features that are the patterns in how they connect with friends across the entire social network including things like the age or gender distribution of their friends. More than 20,000 deep features are used for this.

2.5 Research Papers

[1] proposed a system consisting of these parts:

- A decentralized publishers management protocol to assess the authencity of news outlets.
- Smart contract to publish articles along with its metadata i.e., publisher, date and time, public key and so on.
- A blockchain with honest miner nodes to prevent nefarious agents from tampering.

[8] goes even deeper. Several concepts are proposed that can be applied in the scope of journalism which include blockchain and other distributed ledger technologies (DLTs).

Blockchain can be implemented as an incentivized truth system, in which reliable truth-seekers get incentives as their reputation grows. 4Facts.org, a now defunct platform is cited by the author. The challenges noted by the author are user biasness, malicious actors, proof of

authenticity before entry of data in blockchain. Another disadvantage is that content moderation is less clear in a blockchain-a decentralized network.

[17] detects misinformation by reviewing its characteristics and its disclosure. In addition to using Naïve bayes which gives a prediction precision between 70-76%, they also used POS textual analysis.
[4] uses keywords to extract real-time tweets and extracts important features from it after preprocessing them.

- The first algorithm, Logistic Regression was able to achieve 93.8% accuracy which is almost too perfect.
- The second algorithm, Naïve Bayes gave only 72.5% due zero frequency.
- The third algorithm, Long Short-Term Memory, gave 50.5% after 5 epochs.
- The last algorithm, Support Vector Machine gave a 92.5% accuracy.

Chapter 3: Design and Development

3.1 Overall Description

3.1.1 Product Perspective

DigiAuth aims to reduce the spread of misleading information and rumors and to empower the readers to make a more fact-based judgment about what they see online.
It works like two-way authentication. First, AI algorithms are applied to the data, after that our data is imputed into a blockchain ledger where further it is processed for its authenticity.

3.1.2 Product Functions

The main function is to detect whether the news is fake or not from a given text of input data.

3.1.3 User Classes and Characteristics

3.1.3.1 Summary of User Classes

The following section describes the types of users of the DIGIAUTH. There are explanations of the user followed by the interactions the user(s) shall be able to make with the software.

Typical Users

Typical Users include the general public that is interested in knowing the authenticity of the news.

Programmers

This project is perfect for programmers who are interested in AI or blockchain and want to make improvements to the current project.

Journalism Purposes

This project is also for those people who work in the journalism industry and want to convey REAL news through their channels.

Social Media Purposes

Instead of spreading fake news and instigating confusion and commotion in public, authentic news can be shared.

3.1.4 Operating Environment

The software will work for Windows 7 and above and for all Linux distributions.

3.1.6 Assumptions and Dependencies

- The data inputted is only of the text field.
- The input data is already pre-processed.

3.2 External Interface Requirements

3.2.1 User Interfaces

Users will be able to authenticate the news using a fake news detector.

3.2.2 Hardware Interfaces

No hardware interface is required.

3.2.3 Software Interfaces

DigiAuth provides a platform for a software interface.

A web application

3.2.4 Communications Interfaces

Not yet specified till this version of SRS.

3.3 System Features

This section illustrates organizing the functional requirements for the project DIGIAUTH by system features.

3.3.1 Functional Requirements

The description and priority table is shown below.

S/No.	Module	Priority (5 for highest 1 for lowest)
1	News Collection from datasets	4
2	Authentication using Authenticators	5
3	Authentication using AI	4
4	Uploading of Authenticated news on a ledger	4
5	Rating of the sources	3
6	Maintaining the ledger	3

Figure 2- Description and Priority Table of Reqs

The Requirements Detail are given below:

REQ-1: Solidity for Backend of Blockchain

REQ-2: Fetch News from source and dump into sink

REQ-3: Apply ML algorithms

REQ-4: Register authenticators

REQ-5: Input in Ledger

REQ-6: Analyze News

REQ-7: Provide reputation scores

REQ-8: Show news to general public

REQ-9: Internal authenticators display dashboard

Below are the Use Case Specifications for this project.

Use Case ID	1
Use Case	Fetch News
Primary Actor	News Sources
Description	To fetch the news from news sources and dump it into the sink so that we can visit
Pre-Condition	Cron Job
Normal Flow	News is fetched and sent to sink
Alternate Flow	If there are any exceptions, the news will not be loaded or will not go to sink.
Post-Condition	News is fetched and into a sink

Use Case ID	2
Use Case	Apply algorithms
Primary Actor	Agent
Description	ML techniques for deep learning and transfer learning are applied.
Pre-Condition	Pre-processed data
Normal Flow	the algorithm will flag the news as fake or genuine or alternatively
Alternate Flow	There is an exception in the dataset or the flow of detection.

Post-Condition	The News is flagged.

Use Case ID	3
Use Case	Register Authenticators
Primary Actor	User/Admin/Authenticator
Description	Authenticators are registered
Pre-Condition	Valid information for the registration of the user
Normal Flow	The user gets registered in the system and information is stored in the database.
Alternate Flow	Invalid arguments or information in the registration, user registration failed.
Post-Condition	The user is registered and now can access the system.

Use Case ID	4
Use Case	Input in ledger
Primary Actor	User, Agent
Description	Ledger for all information
Pre-Condition	News to be classified by the Authenticators

Normal Flow	News is classified and inserted into the ledger
Alternate Flow	There is an exception in consensus and the transaction failed.
Post-Condition	News classified is in the ledger for the people to see.

Use Case ID	5
Use Case	Analyze News
Primary Actor	ML Model Agent, Authenticators
Description	Authenticators to analyze the news
Pre-Condition	News sink
Normal Flow	The parameters are classified, and news is flagged.
Alternate Flow	Invalid news or exception
Post-Condition	News is classified.

Use Case ID	6
Use Case	Provide Reputation Scores
Primary Actor	Consensus algorithm
Description	Providing scores to authenticators
Pre-Condition	Scores provided to the news by different authenticators

Normal Flow	An algorithm increments or decrements the reputation score to the authenticator
Alternate Flow	No valid authenticator
Post-Condition	Reputation for the authenticators are set

Use Case ID	7
Use Case	Show news
Primary Actor	Users, Public
Description	Flagged news can be accessed by the general public
Pre-Condition	Directory of flagged news
Normal Flow	Dashboard for the directory is accessible to the users for view
Alternate Flow	Server error, not available
Post-Condition	News is accessible

Use Case ID	8
Use Case	Internal/Authenticators Display dashboard
Primary Actor	User

Description	Using Web3Js application
Pre-Condition	User must have knowledge credentials
Normal Flow	The news sources and list is available to analyze and flag
Alternate Flow	No news is available
Post-Condition	Dashboard is displayed

3.4 Nonfunctional Requirements

3.4.1 Performance Requirements

Performance is vital in projects that require real-time computations. This statement is also true for our project. A critical aspect of our project is to authenticate the news and input data.

3.4.2 Safety Requirements

An Unauthorised user can not access information

Information should not be leaked.

Information should not be distorted.

3.4.3 Security Requirements

Where the data is stored, it should not be destroyed.

No other user can be on the server of the application.

Have backup.

3.4.4 Software Quality Attributes

Flexibility - The system can be adapted for different environments and circumstances to deal with the changes in the business environment.

Availability - The system is responsible for the ability to continue to operate under predefined conditions

3.4.5 Business Rules

The system is available for Linux and Windows only.

DigiAuth solution is suitable for small networks.

3.5 System Overview

The main idea behind the project is to create a system to flag news as fake or real. The system will take *text* data by scrapping the online news sites(predefined) and after preprocessing the data, NLP techniques would be applied on it to classify it as fake or true. The next module will be about registering the authenticators into the blockchain ledger. The ML classified news will be inputted into the ledger where the authenticators will authenticate it and flag it as fake or true. This flagged news will be visible to the public through the web screen along with the reputation score of the authenticator provided by a consensus algorithm.

3.6 System Architecture

3.6.1 Architectural Design

Overview of Modules/ Components

Provided are the modules that our program will be structured upon.

Acquisition of Dataset

A cronjob will scrape the text data from specific online news sites and form a dataset.

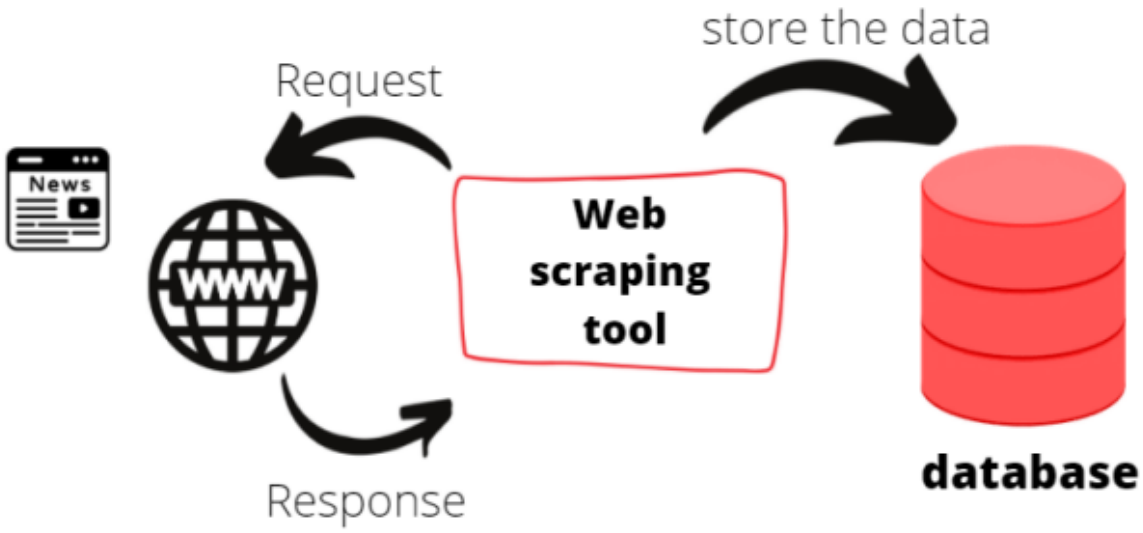


Figure 3- Acquisition of dataset

Application of NLP techniques on dataset

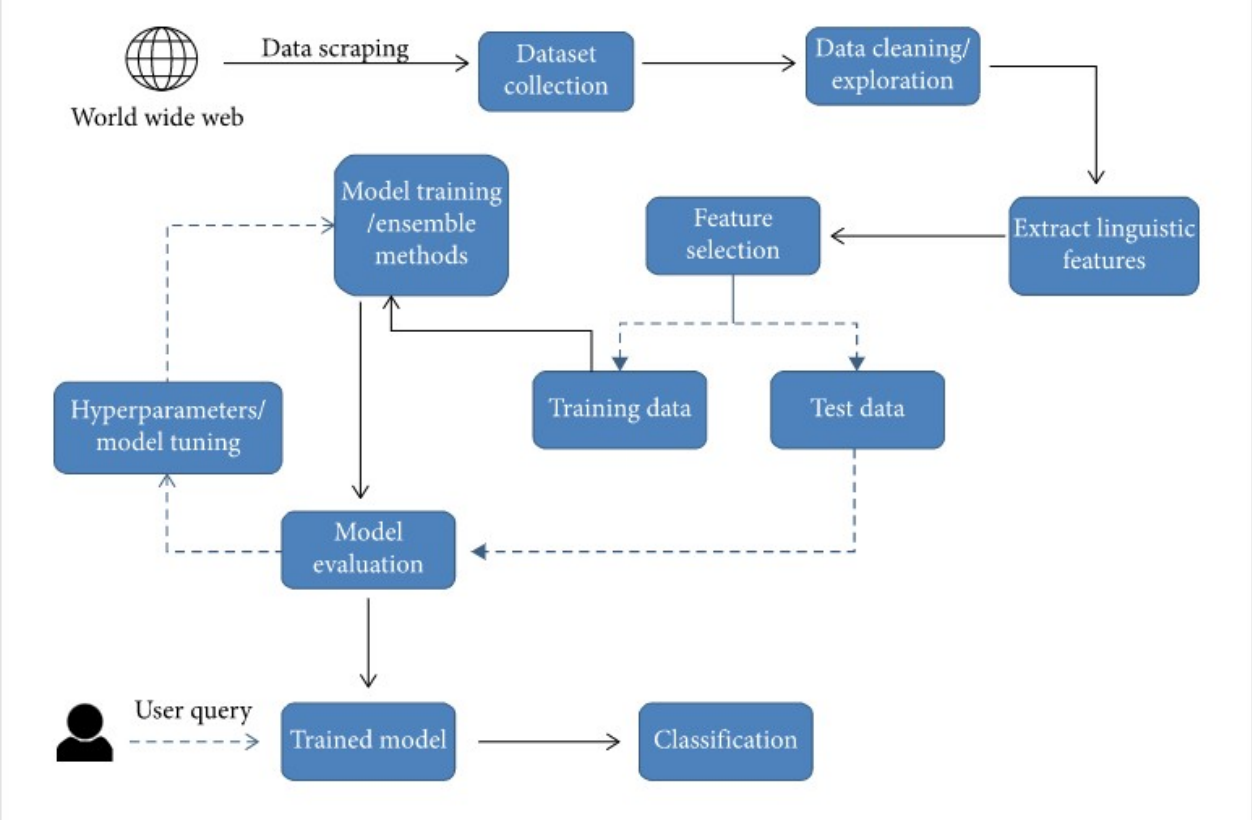


Figure 4- Application of NLP techniques

Blockchain ledger

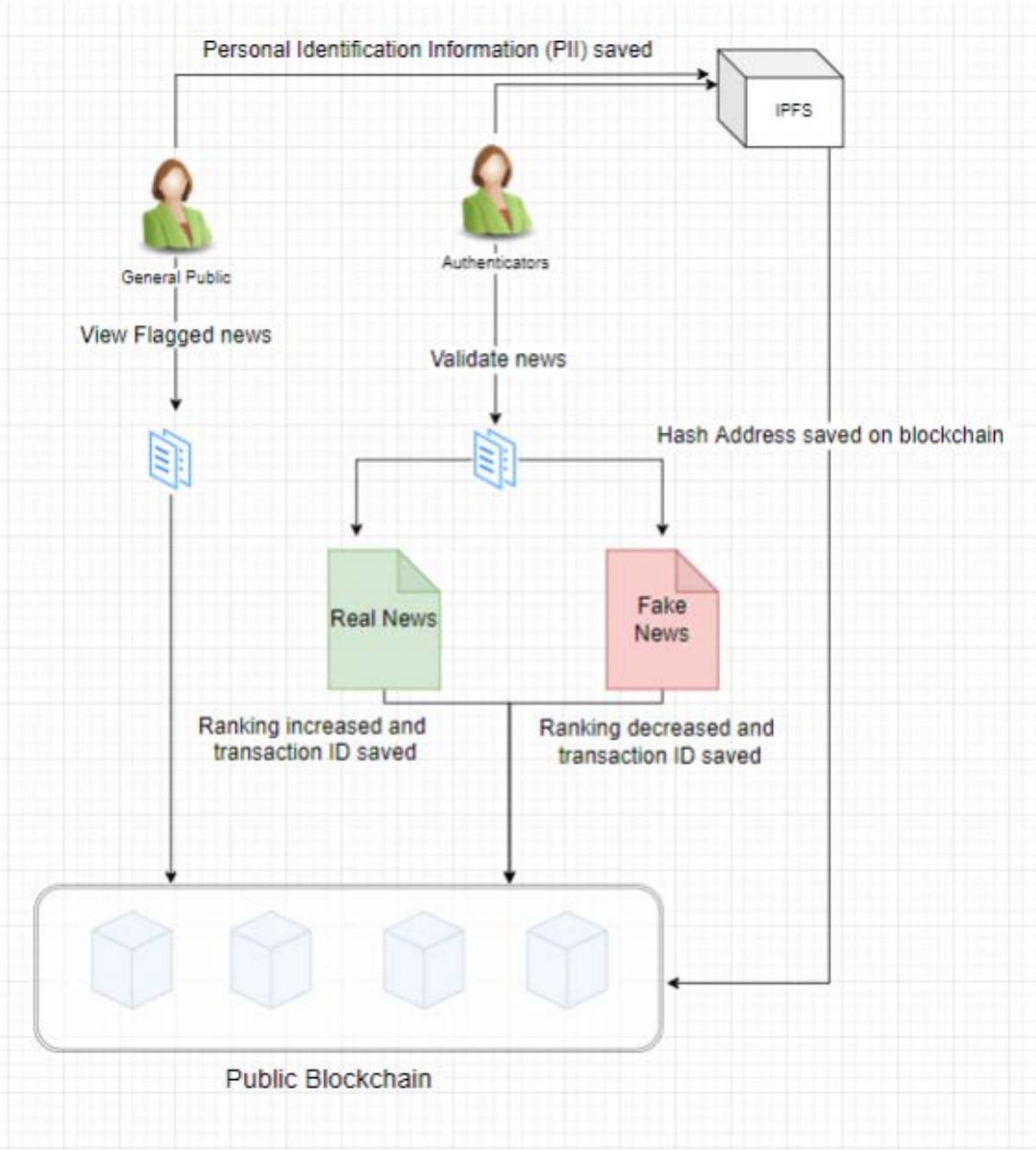


Figure 5- Blockchain Ledger

Register Authenticators and Provide them Scores

Authenticators would be registered, and they will authenticate the news. Based on that they will be given scores.

Web Application Interface

A web application will be made where the general public can see the flagged news.

3.6.2 Decomposition Diagram

The diagram(s) show the higher-level description of the application(s), generic working of the application(s) and interaction with the user.

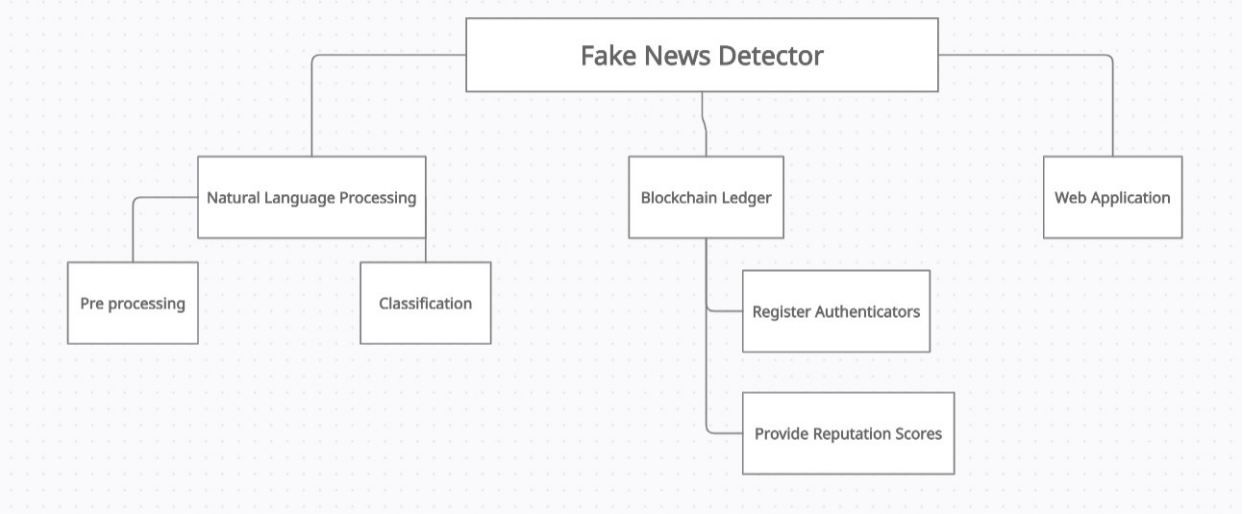


Figure 6- Decomposition Diagram

3.6.3 Design Rationale

Our main architectural pattern followed will be a **data flow architecture** focusing more on the **DeHiDe** architecture for the blockchain part.

This framework filters out the news that is real from what is not real by combining a deep learning model and blockchain.

Following data flow design will be used consisting of the following features:

- Web scraping to acquire data from online news sites into database
- Preprocessing data to remove date, title, punctuation, remove stop words and convert into lowercase and then classifying the data into fake or true.
- Inputting the data into the ledger where authenticators are already registered, they will flag the news.
- Web application where the flagged news is visible.

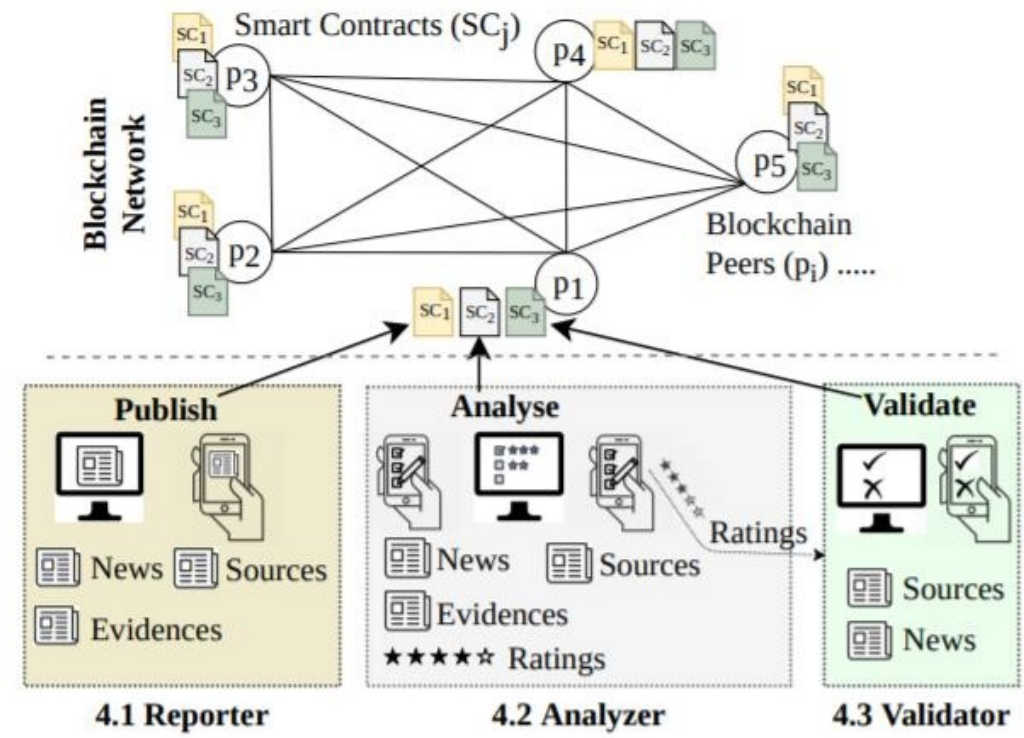


Figure 7- Architectural Overview

The flow of data is shown in the figure below,

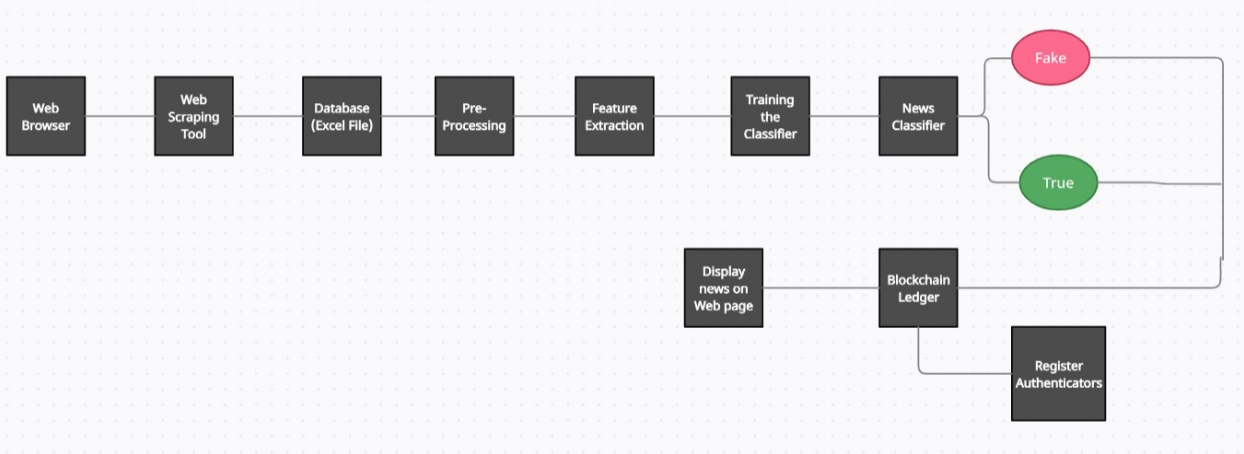


Figure 8- Data Flow Diagram

User Interface Issues

To begin the process, we only allow the authenticators to validate the news predefined, no user can insert any news by themselves, no modification/alteration is allowed in the system. After the score is added to the news by the users it will be displayed in the directory/dashboard for public to see and use it as a reference. The news itself is in textual form and no other data format is supported as of now.

3.6.4 Data Design

Data Description

The information domain of our system will be passed form the backend database of the program saved when the data is synced from the predefined sources in a scheduled task. That textual data is processed and transformed into the labels and features so that Machine Learning can be applied.
The basic database object handling will be used to the CRUD operations and preservation of the data modules.

Data Dictionary

Basically, the data on which we perform operations is mostly text and string in specific, all the stop words, punctions, dates, and titles are removed to fit the confusion matrix. Initially when the module Artifactor is started it creates the database with the news, this runs periodically to sync the news in the form of scheduled task.

3.6.5 Component Design

Component Architecture is the “context” of the system’s use.

3.6.5.1 Overview of Modules/Components

Input Module

The input component will take text data from a web scraping tool and convert it into a presentable format on which Machine Learning algorithms would be applied. Text would be analyzed and passed over to the Blockchain component.

Analysis of Text:

This analysis on the news text is done on the following perspectives:

1. Computational Linguistic

    a. Question Mark, Linguistic, Exclamations and Capital Letters

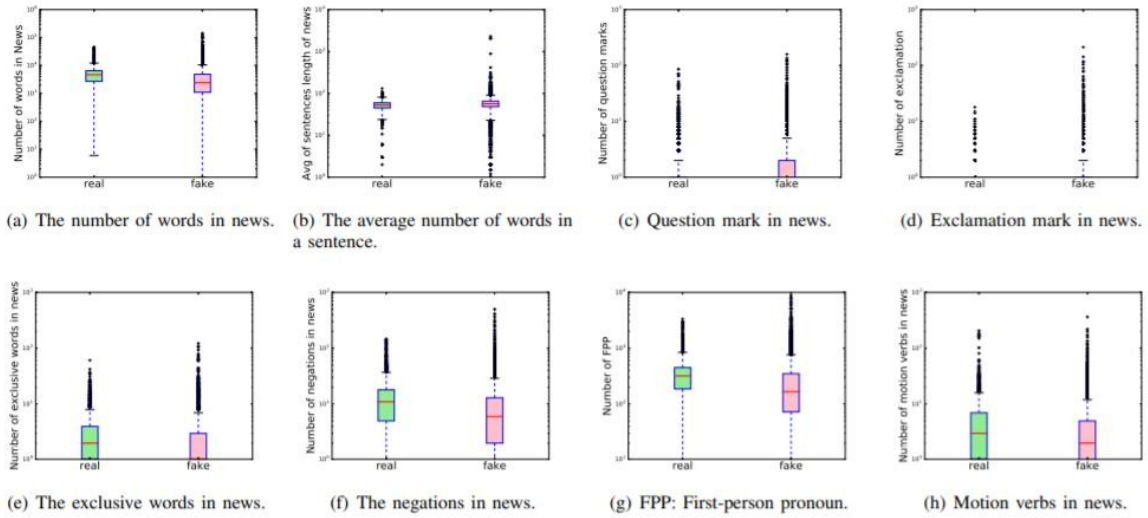
    Statistics show that real news has fewer question marks due to lack of rhetorical questions emphasizing in the ideas and sentiments in real news.
    On the other hand, both have fewer exclamation marks as shown in Fig-9. But Fake has a bigger box plot, due to the use of emotional outburst in even declarative statements.
    Fake also uses more capital letters to draw attention of users and enticing them to believe it.

    b. No. of Words in Sentences

    Liars do have some control over what they are saying, but their mind may leak out through whatever style they are using to tell their story. The same thing applies to liars writing rumors and fake news. Figure 9 shows that the variance in real is small as compared to variance in fake news regarding the no. of words and sentences as the real news is written under several restrictions on the editor including grammar, no. of words, length of sentence, etc.

c. Cognitive Perspective

Words e.g., ‘but’, ‘without’, ‘however’ and negations e.g., ‘no’, ‘not’ are used more frequently in true news. The same thing happens with negations. Fake does not use them to lower the probability of getting caught from contradictions.



(a) The number of words in news. (b) The average number of words in a sentence. (c) Question mark in news. (d) Exclamation mark in news.

(e) The exclusive words in news. (f) The negations in news. (g) FPP: First-person pronoun. (h) Motion verbs in news.

Figure 9-Computational Linguistic Difference in fake and true news

2. Lexical Diversity

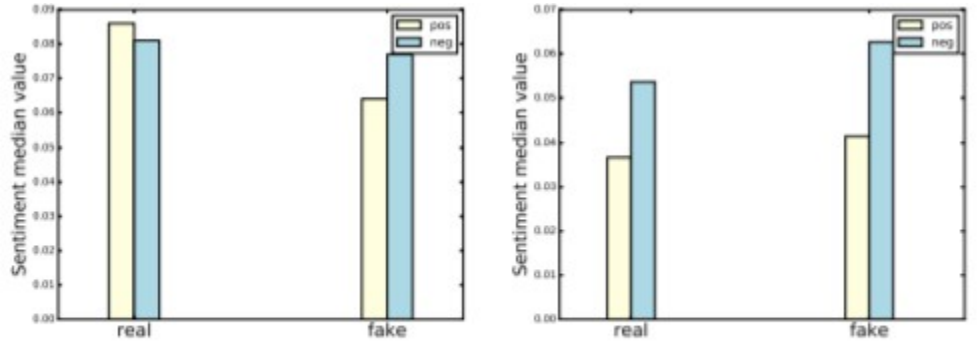
This measures the different words used in a text, while lexical density measures proportion of lexical items (e.g., nouns, verbs, adjectives, and some adverbs) in the text. Real news is more diverse. The lexical diversity of real news is 2.2e-06, which is larger than 1.76e-06 for fake news

3. Psychology Perspective

Deceptive people often use language that do not give any reference to themselves. A liar does not use words like “I” and “we”. Instead of saying “I didn’t take your book,” a liar might say “That’s not the kind of thing that anyone with integrity would do”. Fake news has less frequency of usage of first-person and third-person pronouns. Deception uses less first and third person pronouns.

4. Sentiment Analysis

Real news is more positive than fake news. The deceivers under the tension and guilt and not being confident in the topic, have more negative than positive emotions. The standard deviation is also higher in fake news showing the string sentiments it has.



(a) The median sentiment values: positive and negative.

(b) The standard deviation sentiment values: positive and negative.

Figure 10-Semantic Analysis to analyze news

Blockchain Component

The data from the input component will be inputted into the ledger where the authenticators are already registered. They will analyze the news and flag it accordingly. The flagged news will go to the output component along with the reputation scores of the authenticators.

Output Component

A web application will display the flagged news along with the reputation scores to the general public.

3.6.5.2 Component Diagram

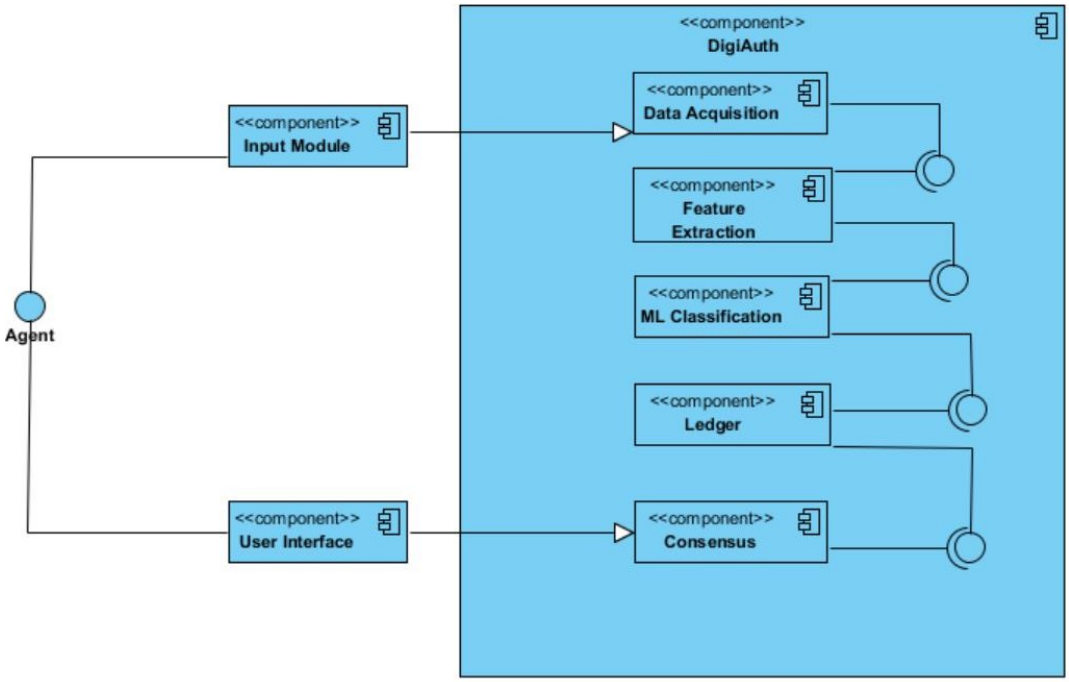


Figure 11-Component Diagram

3.6.5.3 Use Case Diagram

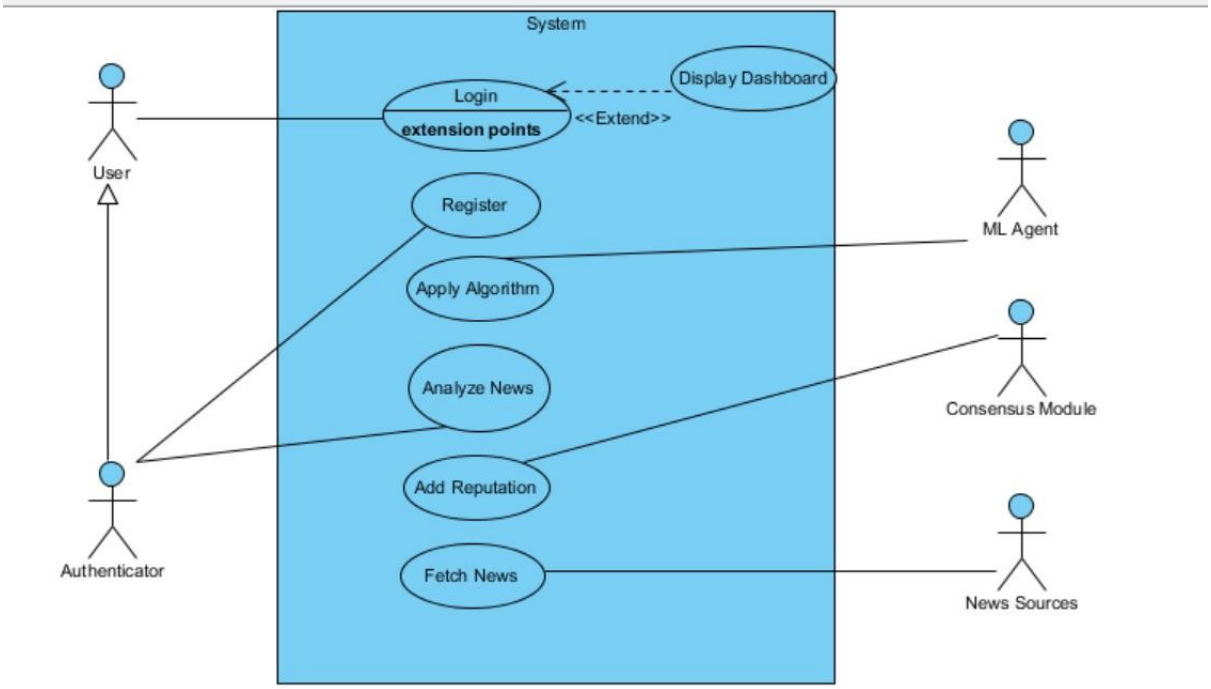


Figure 12-Use Case Diagram

3.6.5.4 Activity Diagram

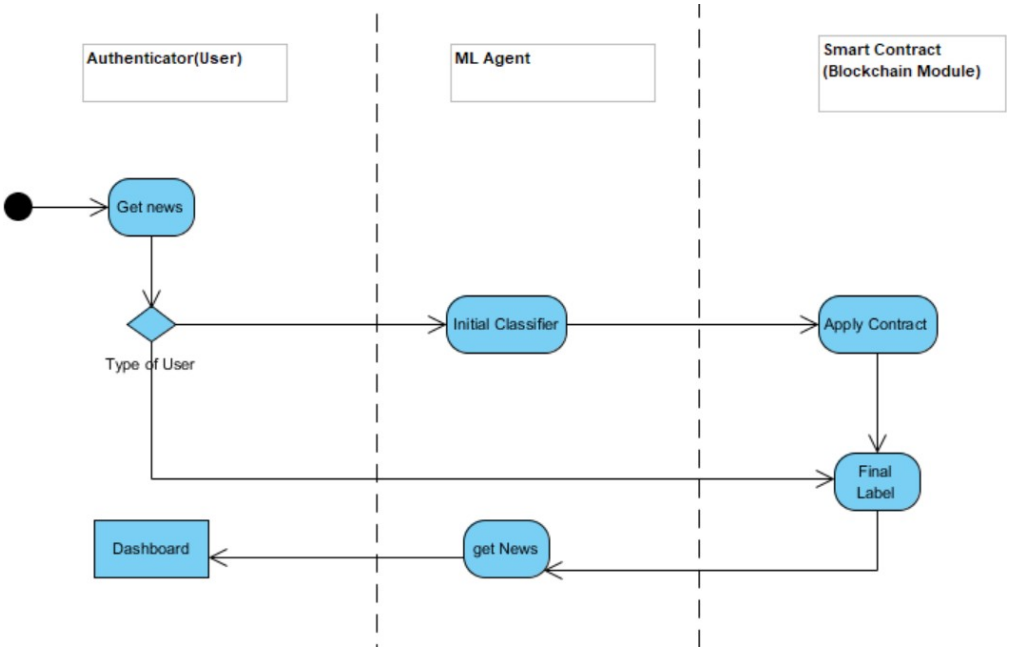


Figure 13- Activity Diagram

3.6.5.5 Sequence Diagram

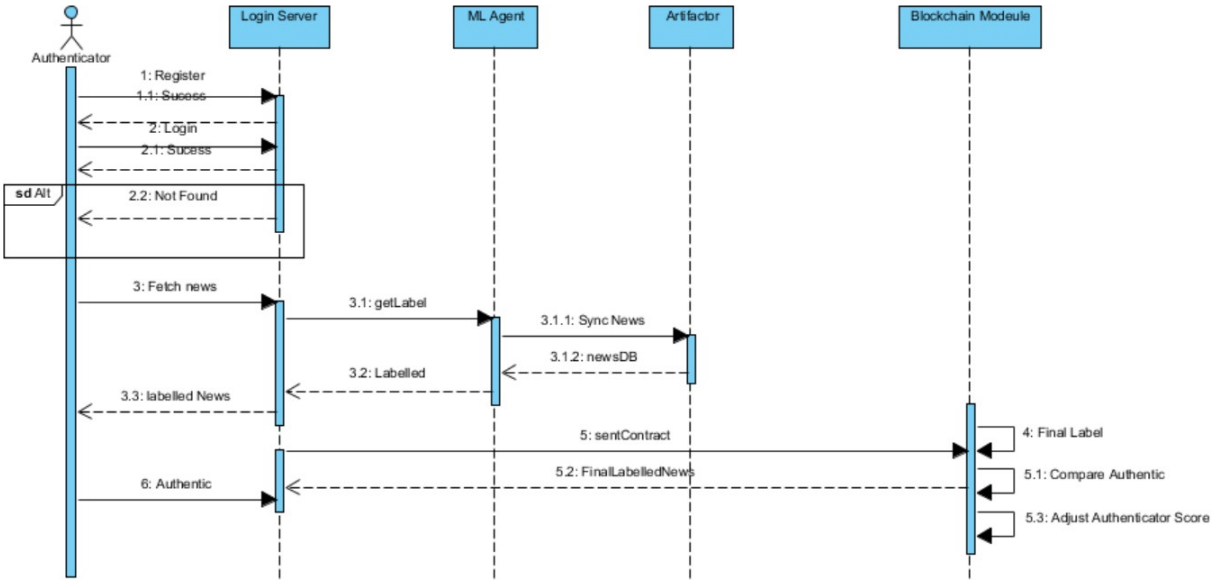


Figure 14- Sequence Diagram

3.6.5.6 Class Diagram

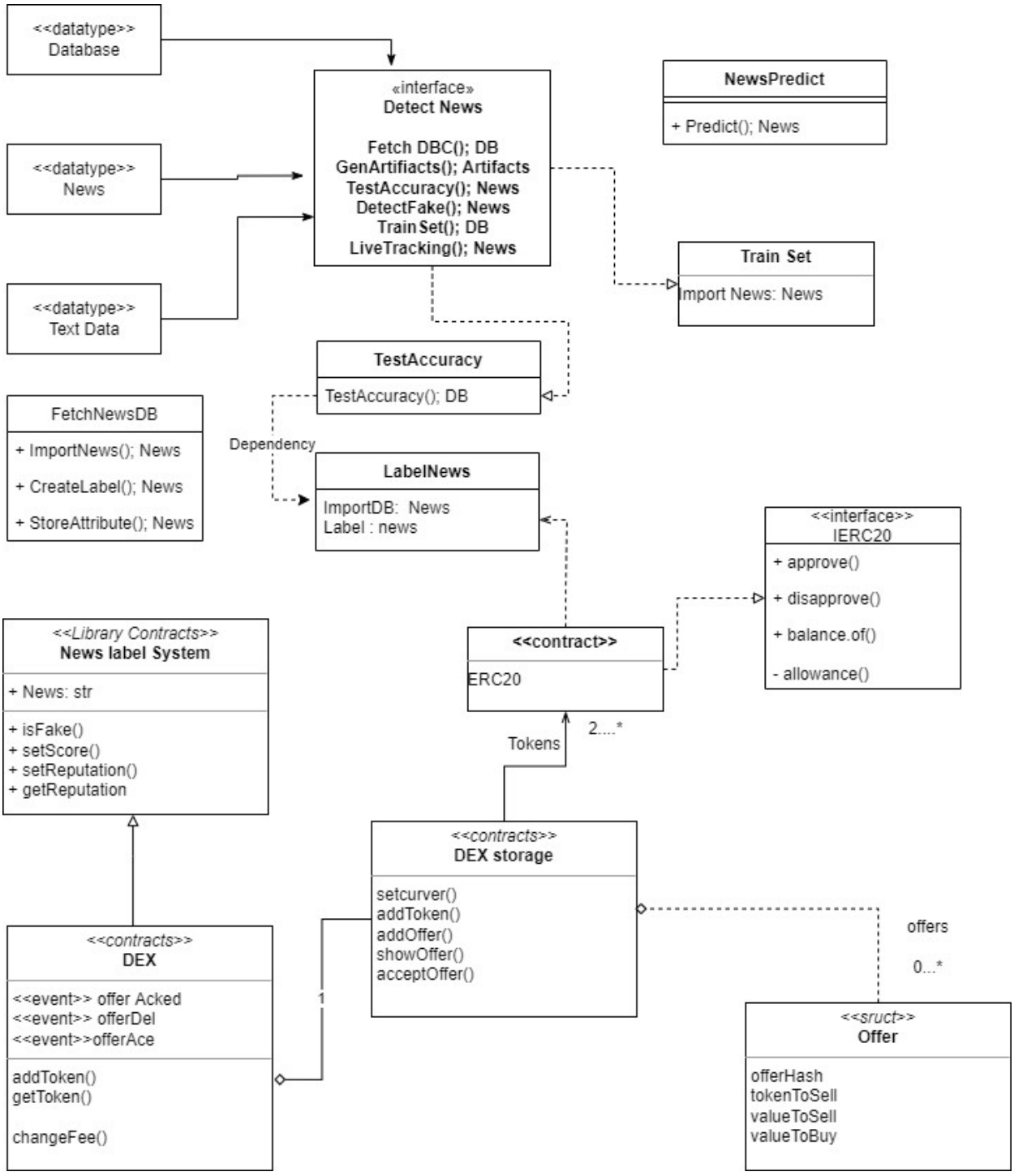


Figure 15- Class Diagram

Chapter 4: Code Analysis and Evaluation

4.1. Blockchain

4.1.1. Solidity

It is being used to develop the contacts, following are the characteristics of the solidity language.

- Solidity is the statically typed language used to create smart contracts and generate a chain of transaction records in the blockchain system.
- Contracts Inherit all Members from Address.
- Specify value and gas for function calls
- A tool for creating machine-level code and then compiling the code on the Ethereum Virtual Machine (EVM).
- Similar to C and C++. For example, a “main” in C is equivalent to a “contract” in Solidity.

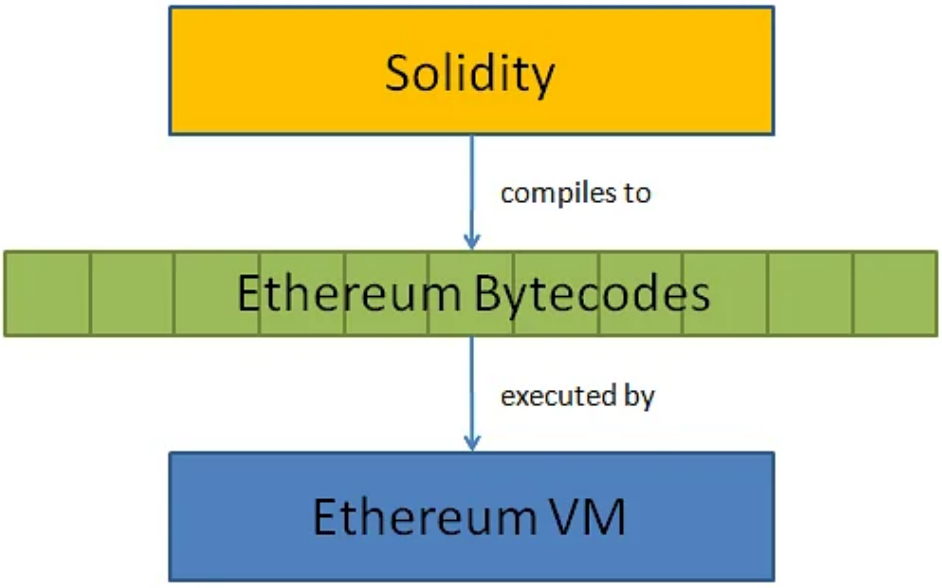


Figure 16- Solidity Diagram

4.1.2. Smart Contracts

These are high-level codes that are compiled into the Ethereum Virtual Machine before they are executed in the blockchain. They enable users in trustworthy transactions that are traceable and cannot be reversed without the interference of any third-party agent. Serpent, Solidity, LL and Mutan are some of the programming languages that are used to write these smart contracts. For this project we will use solidity which allows us to execute in both online and offline modes.

4.1.3. Requirements for Offline Mode

For an offline mode, three softwares are required.

- Node.js
- Truffle globally.
- Ganache-cli.

4.1.4. User Interface

We use React which is a JavaScript-based UI development library because of its ability to easily create dynamic applications that require less coding with greater functionality. It's use of Virtual DOM also improves performance by creating web applications faster.

Figure 17- Dynamic Application Creation

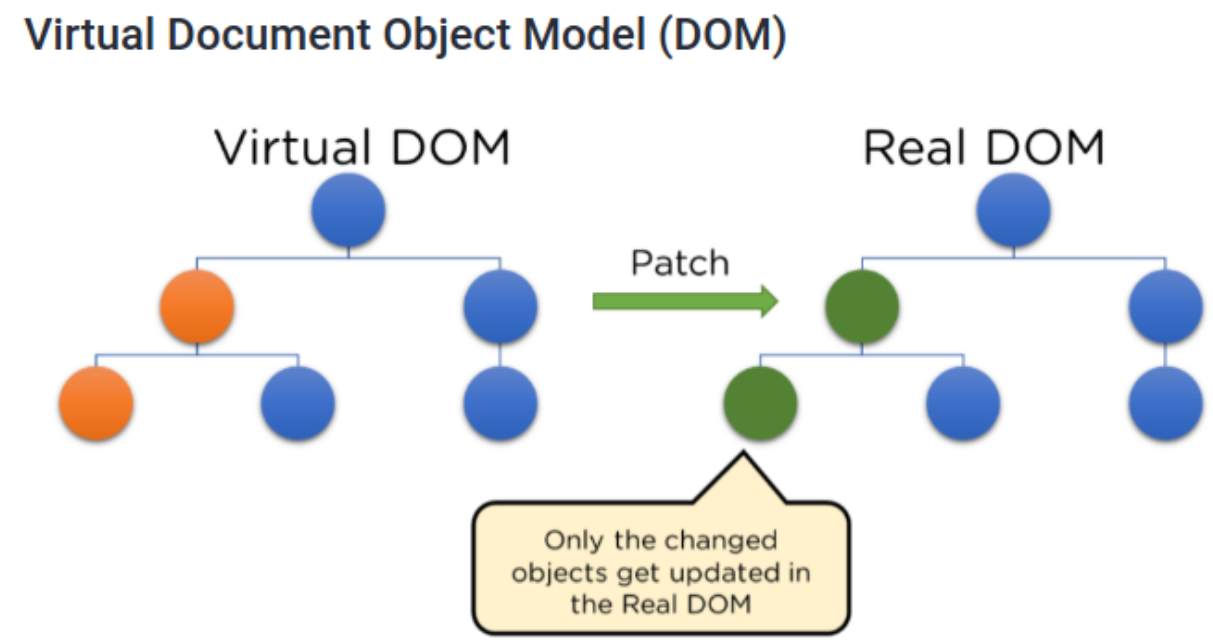


Figure 18- Virtual Document Object Model

4.3 Machine Learning

We used the following datasets to train the model

1. https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
2. https://www.kaggle.com/datasets/jruvika/fake-news-detection

The following algorithms were used:

1. Logistic Regression
2. Decision Tree Classifier
3. Gradient Boosting Classifier
4. Random Forest Classifier
5. Naïve Bayes Algorithm
6. Long Short-Term Memory (LSTM)
7. Passive Aggressive Classifier

4.3.1 Logistic Regression

Logistic Regression uses the sigmoid function to map any real value between 0 and 1.

$$log\left(\frac{P}{1-p}\right)$$

	precision	recall	f1-score	support
0	0.98	1.00	0.99	5853
1	0.99	0.82	0.89	550
accuracy			0.98	6403
macro avg	0.99	0.91	0.94	6403
weighted avg	0.98	0.98	0.98	6403

Figure 19- Classification Table (Logistic Regression)

4.3.2 Decision Tree Classifier

It builds classification models in a tree like structure by breaking down a dataset into smaller and smaller subsets.

$$Gain(S, A) = H(s)\frac{|S_v|}{|s|}H(S_v)$$

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5853
1	0.99	0.96	0.98	550
accuracy			1.00	6403
macro avg	0.99	0.98	0.99	6403
weighted avg	1.00	1.00	1.00	6403

Figure 20- Classification Table (Decision Tree)

4.3.3 Gradient Boosting Classifier

It minimizes the prediction error by combining the next model with the previous ones in order to come to a best possible model.

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + Const$$

	precision	recall	f1-score	support
0	1.00	1.00	1.00	5853
1	0.98	0.99	0.99	550
accuracy			1.00	6403
macro avg	0.99	0.99	0.99	6403
weighted avg	1.00	1.00	1.00	6403

Figure 21- Classification Table (Gradient Boosting)

4.3.4 Random Forest Classifier

It consists of a forest of many decision trees and overall the prediction is more than it can be of an individual tree.

	precision	recall	f1-score	support
0	0.96	1.00	0.98	5853
1	1.00	0.59	0.75	550
accuracy			0.97	6403
macro avg	0.98	0.80	0.86	6403
weighted avg	0.97	0.97	0.96	6403

Figure 22- Classification Table (Random Forest)

4.3.5 Naïve Bayes Algorithm

It uses conditional probability to make predictions. The simple equation for Bayes theorem is given below

$$P(A|B) = P(B|A) * \frac{P(A)}{P(B)}$$

	precision	recall	f1-score	support
0	0.95	0.96	0.95	5849
1	0.48	0.42	0.45	554
accuracy			0.91	6403
macro avg	0.72	0.69	0.70	6403
weighted avg	0.91	0.91	0.91	6403

Figure 23- Classification Table (Naive Bayes)

4.3.6 Long Short-Term Memory

They are recurrent neural networks that can learn order dependence in sequence prediction problems.

$$i_t = \sigma\left(w_i\left[h_{t+J}x_t\right] + b_i\right)$$
$$f_t = \sigma\left(w_f\left[h_{t+J}x_t\right] + b_f\right)$$
$$o_t = \sigma\left(w_o\left[h_{t+J}x_t\right] + b_o\right)$$

	precision	recall	f1-score	support
0	0.99	1.00	0.99	5881
1	0.99	0.99	0.99	5344
accuracy			0.99	11225
macro avg	0.99	0.99	0.99	11225
weighted avg	0.99	0.99	0.99	11225

Figure 24- Classification Table (LSTM)

4.3.7 Passive Aggressive Algorithm

They do not require a learning rate similar to a Perceptron model, but they do need a regularization parameter.

They are called Passive Aggressive because they remain passive (do not change) in case of correct prediction and become aggressive (make changes to model) if prediction is incorrect.

	precision	recall	f1-score	support
FAKE	0.93	0.93	0.93	615
REAL	0.93	0.94	0.93	652
accuracy			0.93	1267
macro avg	0.93	0.93	0.93	1267
weighted avg	0.93	0.93	0.93	1267

Figure 25-Classification Table(Passive Aggressive)

Chapter 5: Conclusion and Future Works

Classification of any type of news requires an in-depth knowledge of the domain and requires expertise in the field if anomaly identification of textual data. The dataset that we used for this project is collected from the entire Internet and has news of various other domains other than the political ones.

Some models that we used to classify show better results as compared to others but in short the results were similar due to the extensive data used.

The accuracy of each model is mentioned below:

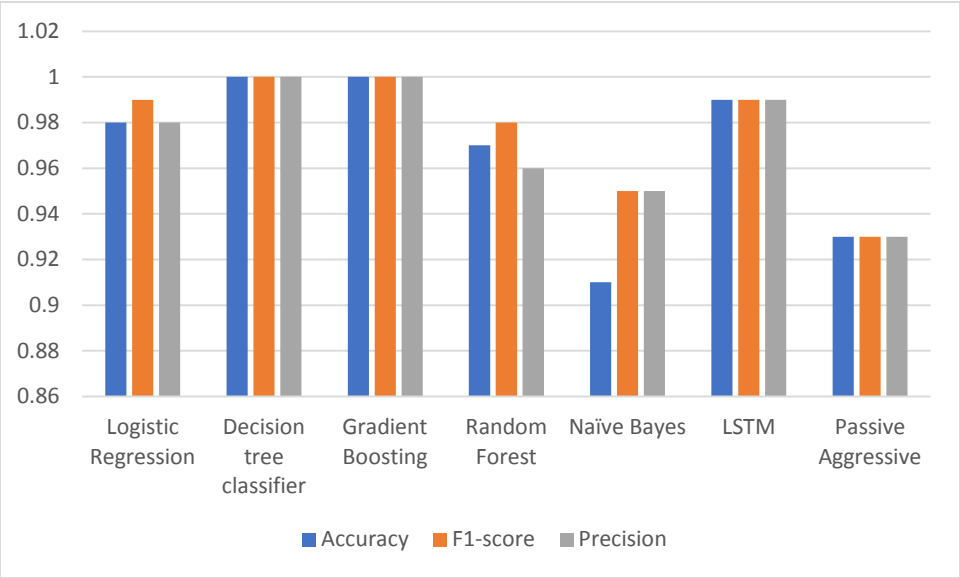


Figure 26- Comparison chart for models

In addition to Machine learning, Blockchain gives a tamper proof system.

This project serves as a two-factor authentication system of the news: firstly, using NLP and classifying news based on historical data, secondly, using p2p Ethereum to register news and having reputable authenticators giving its proof.

What we aimed from this project is to explore and delved deeper into the difference between what is real and what is not. We analyzed text from the World Wide Web in hopes of finding a pattern that could classify what is true and what is not. From a lexical, to linguistic to psychological perspective, we explored the line between “fake” and “real” news easily understandable in a mathematical perspective.

The use of Blockchain was something new for us. Using this decentralized system, we built a tamperproof registry. But unfortunately for this registry to work we need reputable authenticators in the form of experienced journalists, reporters, etc. which up to this date we do not have.

If this project is funded in the future or a startup is made on this basis, we hope to get authenticators for this blockchain project.

Another thing for future works is to make the model learnable so that it can retrain itself with new data. Also, to make our web scraper automatic and build a data register to store all the data that is being scraped up to this day and not only the current day’s data which is what our web scraper is doing right now.

Chapter 6: References and Citations

[1]

Qayyum, Adnan & Qadir, Junaid & Janjua, Muhammad & Vira, Falak Sher. (2019). Using Blockchain to Rein in the New Post-Truth World and Check the Spread of Fake News. IT Professional. 21. 16-24. 10.1109/MITP.2019.2910503.

[2]

Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, Muhammad Ovais Ahmad, "Fake News Detection Using Machine Learning Ensemble Methods", Complexity, vol. 2020, Article ID 8885861, 11 pages, 2020. https://doi.org/10.1155/2020/8885861

[3]

Yang, Yang & Zheng, Lei & Jiawei, Zhang & Cui, Qingcai & Li, Zhoujun & Yu, Philip. (2018). TI-CNN: Convolutional Neural Networks for Fake News Detection.

[4]

Ahmed, Alim Al Ayub & Aljabouh, Ayman & Donepudi, Praveen & Choi, Myung. (2021). Detecting Fake News Using Machine Learning: A Systematic Literature Review.

[5]

Tavishee Chauhan, Hemant Palivela. Optimization and improvement of fake news detection using deep learning approaches for societal benefit, International Journal of Information Management Data Insights, Volume 1, Issue 2, 2021, 100051, ISSN 2667-0968 https://doi.org/10.1016/j.jjimei.2021.100051

[6]

Z. Shae and J. Tsai, "AI Blockchain Platform for Trusting News," 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), 2019, pp. 1610-1619, doi: 10.1109/ICDCS.2019.00160.

[7]
Akshada Babar, Nalini Jagtap, Akshata Mithari, Aakash Shukla and Prachi Chaudhari. A Survey on Fake News Detection Techniques and using a Blockchain based System to Combat Fake News. *International Journal of Computer Applications* 176(27):47-53, June 2020.

[8]
P. Fraga-Lamas and T. M. Fernández-Caramés, "Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality," in IT Professional, vol. 22, no. 2, pp. 53-59, 1 March-April 2020, doi: 10.1109/MITP.2020.2977589.

[9]
Mohamed Torky, Emad Nabil and Wael Said, “Proof of Credibility: A Blockchain Approach for Detecting and Blocking Fake News in Social Networks” International Journal of Advanced Computer Science and Applications(IJACSA), 10(12),
2019. http://dx.doi.org/10.14569/IJACSA.2019.0101243

[10]
2019 7th International Conference on Smart Computing and Communications (ICSCC), Shovon Paul, Jubair Islam Joy, Shaila Sarker. Fake news detection in social media using Blockchain

[11]
Deeptrustalliance.org

[12]
The News Provenance Project

[13]
2019 published in The Innovation Machine, George Krasadakis. Solving ‘Fake news’, with AI, Blockchain and a Global Community

[14]
The Social Truth, an open distributed digital content verification for hyper-connected sociality

[16]
Forbes Technology Council, Chief Evangelist at Rockmetric. CXO Advisor, Author, Startup Mentor, TEDx Speaker and Dogfather. Build The Truth Block by Block

[17]
Z Khanam *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1099** 012040