A PROJECT REPORT

ON

"INTELLIGENT NEWS AGGREGATOR AND VALIDATOR"

Submitted to

UNIVERSITY OF MUMBAI

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER ENGINEERING

BY

KHAN SHAHID JAHID SAIRA	**12CO40**
PRASAD CHANDRESH KAMLESHWAR MANTI	**11CO35**
KHATRI AMAANULLA ABDUL SATTAR RESHMA	**13CO45**
KHAN IRFAN ABUBAKAR MATHARUNNISHA	**13CO36**

UNDER THE GUIDANCE OF
PROF. TABREZ KHAN

DEPARTMENT OF COMPUTER ENGINEERING
Anjuman-I-Islam's Kalsekar Technical Campus
SCHOOL OF ENGINEERING & TECHNOLOGY
Plot No. 2 & 3, Sector - 16, Near Thana Naka,
Khandagaon, New Panvel - 410206

2018-2019

AFFILIATED TO

UNIVERSITY OF MUMBAI

A PROJECT II REPORT
ON

"INTELLIGENT NEWS AGGREGATOR AND VALIDATOR"

Submitted to
UNIVERSITY OF MUMBAI

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN
COMPUTER ENGINEERING

BY

KHAN SHAHID JAHID SAIRA	**12CO40**
PRASAD CHANDRESH KAMLESHWAR MANTI	**11CO35**
KHATRI AMAANULLA ABDUL SATTAR RESHMA	**13CO45**
KHAN IRFAN ABUBAKAR MATHARUNNISHA	**13CO36**

UNDER THE GUIDANCE OF
PROF. TABREZ KHAN

DEPARTMENT OF COMPUTER ENGINEERING
Anjuman-I-Islam's Kalsekar Technical Campus
SCHOOL OF ENGINEERING & TECHNOLOGY
Plot No. 2 & 3, Sector - 16, Near Thana Naka,
Khandagaon, New Panvel - 410206

2018-2019
AFFILIATED TO

UNIVERSITY OF MUMBAI

Anjuman-I-Islam's Kalsekar Technical Campus

Department of Computer Engineering

SCHOOL OF ENGINEERING & TECHNOLOGY

Plot No. 2 & 3, Sector - 16, Near Thana Naka,

Khandagaon, New Panvel - 410206

CERTIFICATE

This is certify that the project entitled

"INTELLIGENT NEWS AGGREGATOR AND VALIDATOR"

submitted by

KHAN SHAHID JAHID SAIRA	**12CO40**
PRASAD CHANDRESH KAMLESHWAR MANTI	**11CO35**
KHATRI AMAANULLA ABDUL SATTAR RESHMA	**13CO45**
KHAN IRFAN ABUBAKAR MATHARUNNISHA	**13CO36**

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Engineering) at ***Anjuman-I-Islam's Kalsekar Technical Campus, Navi Mumbai*** under the University of MUMBAI. This work is done during year 2018-2019, under our guidance.

Date: / /

(Prof. TABREZ KHAN)
Project Supervisor

(Prof. KALPANA BODKE)
Project Coordinator

(Prof. TABREZ KHAN)
HOD, Computer Department

DR. ABDUL RAZAK HONNUTAGI
Director

External Examiner

Acknowledgements

We would like to take the opportunity to express our sincere thanks to our guide **Prof. TABREZ KHAN**, Assistant Professor, Department of Computer Engineering, AIKTC, School of Engineering, Panvel for his invaluable support and guidance throughout our project research work. Without his kind guidance & support this was not possible.

We are grateful to him for his timely feedback which helped us track and schedule the process effectively. His time, ideas and encouragement that he gave is helped us to complete our project efficiently.

We would like to express deepest appreciation towards **DR. ABDUL RAZAK HONNUTAGI**, Director, AIKTC, Navi Mumbai, **Prof. TABREZ KHAN**, Head of Department of Computer Engineering and **Prof. KALPANA BODKE**, Project Coordinator whose invaluable guidance supported us in completing this project.

At last we must express our sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped us directly or indirectly during this course of work.

KHAN SHAHID JAHID SAIRA
PRASAD CHANDRESH KAMLESHWAR MANTI
KHATRI AMAANULLA ABDUL SATTAR RESHMA
KHAN IRFAN ABUBAKAR MATHARUNNISHA

Project I Approval for Bachelor of Engineering

This project entitled ***Ïntelligent News Aggregator And Validator¨*** by ***Khan Shahid Jahid Saira (12CO40), Prasad Chandresh Kamleshwar Manti (11CO35), Khatri Amaanulla Abdul Sattar Reshma (13CO45), Khan Irfan Abubakar Matharunnisha (13CO36)*** is approved for the degree of ***Bachelor of Engineering in Department of Computer Engineering.***

Examiners

1. .....
2. .....

Supervisors

1. .....
2. .....

Chairman

.....

Declaration

We declare that this written submission represents our ideas in our own words and where others ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

KHAN SHAHID JAHID SAIRA
12CO40

PRASAD CHANDRESH KAMLESHWAR MANTI
11CO35

KHATRI AMAANULLA ABDUL SATTAR RESHMA
13CO45

KHAN IRFAN ABUBAKAR MATHARUNNISHA
13CO36

ABSTRACT

In recent years, the widespread of fake news is growing at an alarming rate. So much so that it has turned into a social issue. So recognizing and distinguishing fake news is an important task which must be accomplished with proper ideas and science behind it. In this project we have implemented an application, which will help users to discern fake news amongst all sort of news which is available on all over internet. The application will provide users with up to date news using simple news aggregation along with news validation, the news will be classified into three categories fake, genuine and neutral. The validation will be performed in two stages; first using machine learning and second based on news feedback provided by the users of our system. Using this two step method we can make a validation system that is dynamic and one that becomes more accurate with time.

Keywords: Machine learning, Fake news

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The news is a source which provides daily information about technology sports, political, information etc. In recent years people most often interact with news using their smart phones, and there are many platforms like Facebook, Google, Twitter etc. which easily provide news as per user interest.

In social media or on Internet not all news are genuine, sometimes the news propagated is made intentionally fake. Fake news, to be precise is a misleading information or a fabricated story about a public figure, political leader or a party or about certain issues. Users consume these fake news without properly assessing the content of the news and fall for the ulterior motive of fake news propagator. This sometimes affect the readers as well as society.

To mitigate the problem of fake news many fake news detection technique and methods were proposed in the past. Many of them works as well with certain percentage of accuracy. Reference [1] uses naive Bayes classifier for detection of fake news, and using this classification technique to detect fake news, the efficiency of 74% was obtained. Reference [4] uses two approach linguistic approach and network approach, in linguistic approach the contents of deceptive messages is extracted and analyzed to recognize the certain language patterns of such deceptiveness and in network approach, network information, such as network queries and structured knowledge are utilized to provide aggregate deception measures. Reference [5] With the help of the News data sets several analyses were conducted to identify linguistic properties that are present in fake news content and fake news detectors gaining accuracy up-to 70% were used. And at-last to be more perspective the accuracy of fake news detection models are compared with human baseline accuracy.

In our proposed system we will be addressing this fake news problem by providing users with a news platform/application from where they can get all their news, along with the validation of the news they are reading.

Our system will provide users every category of news ranging from politics to sports to pop culture to local news. These news will scraped from online news sources. The scraping of news data will be done periodically in small intervals in order to provide users with the latest news all the time. A learning algorithm will be used to provide users with personalized news. Personalized news will be based on

user preferences and viewing history. The user can provide rating of news they are reading as genuine or fake. This rating can be used later on in the validation process.

The news data scraped from the internet will also serve another purpose in our system i.e. to be used by machine learning algorithms to be validated as fake or genuine. This validation information will later be used along with user rating to classify the news with even greater precision.

1.1 Purpose

Along with the rapid development of the World Wide Web, information on Web pages is rapidly inflated and congested with large amounts of news contents. So to overcome the drawbacks which cant detect or stop the escalation of fake news and to identify useful information that satisfies a users interests we have created these project or we can say a strategy. In this project we are going to validate and aggregate news scrapped from web.User will be able to watch/browse news as many as he wants but the catch is there is going to be a validity for the news he watches and he stores. Our strategy that is, the main motto of our project, our system will rate news on the basis of algorithm with extreme preciseness and accuracy. The process of rating will be two fold i.e from mostly fake to mostly genuine. So it will be in users interest, it will be easy for him to differentiate amongst news he is going to surf.

1.2 Project Scope

The proposed project is the Intelligent News Aggregator and Validator. Our system get its input from the internet (in the form of News article) and using classifiers (Naive Bayes classifier) we divide that that article basis of their genuine percentage. After the calculation of all the data it will get stored on database and shown to the user as per their request.

1.3 Project Goals and Objectives

1.3.1 Goals

- Our goal or we must say a milestone, is to protect our users from the source which deliberately publish hoaxes, propaganda and disinformation purporting to be real news.
- And the second most important goal of our project is to examine how this particular method works for this problem given a manually labeled news dataset and to support (or not) the idea of using machine learning for news detection.

1.3.2 Objectives

In our proposed project, Intelligent News Aggregator and Validator we are dealing with fake news problem using current technology such as machine learning as well as artificial intelligence algorithms. This new technology is very effective in terms of classification between fake and genuine news. Some of our objectives are as follows:

- To develop a web application for online news that can aware the people and give them the truth about the political world and worlds whereabouts and happenings.
- It makes use of various technologies to get required crime oriented information more quickly, easily, colorfully and attractively.
- To do this for more widely coverage of distribution and faster dissemination of information and automatically updating of news.
- Anytime, anywhere, anyone can know about the news or information by our application, intuitively our app is free of cost and stress reducer for users who get intimidated by paid applications.

These promising strategy of ours, encourages us to probe further into this issue to eliminate fraud through fake news.

1.4 Organization of Report

The remaining part of the project is organized as follows:

Chapter 2: Presents a review of related work.

Chapter 3: Describes the time management and time utilization during the project implementation.

Chapter 4: Introduces the Software requirement Specification of our project.

Chapter 5: Proposes the project design of the project.It represent the architectural design,front end design and database design of the project.

Chapter 6: Presents implementation details of our project.

Chapter 7: Presents various test cases that are consider.

Chapter 8: Consist various screenshots of project.

Chapter 9: Provides some concluding remarks and direction of our future work.

Chapter 2

Literature Survey

2.1 Fake news detection using naive Bayes classifier

There are several influential articles about automatic deception detection. In the authors provide a general overview of the available techniques for the matter. In the authors describe their method for fake news detection based on the feedback for the specific news in the microblogs. In the authors actually develop two systems for deception detection based on support vector machines and naive Bayes classifier (this method is used in the system described in this paper as well) respectively. They collect the data by means of asking people to directly provide true or false information on several topics – abortion, death penalty and friendship. The accuracy of the detection achieved by the system is around 70%.

This article describes a simple fake news detection method based on one of the artificial intelligence algorithms – Naive Bayes classifier. The goal of the research is to examine how this particular method works for this particular problem given a manually labeled news dataset and to support (or not) the idea of using artificial intelligence for fake news detection. The difference between these article and articles on the similar topics is that in this paper naive Bayes classifier was specifically used for fake news detection; also, the developed system was tested on a relatively new data set, which gave an opportunity to evaluate its performance on a recent data.

In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes theorem with strong (naive) independence assumptions between the features. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable Naive Bayes typically use bag of words features to identify spam e-mail, an approach commonly used in text classification. Naive Bayes classifiers work by correlating the use of tokens (typically words, or sometimes other constructions, syntactic or not), with spam and non-spam e-mails and then using Bayes theorem to calculate a probability

that an email is or is not a spam message.

2.1.1 Advantages of Paper

a. Automatic analytical methods complement and enhance the notoriously poor human ability to discern information from misinformation.

b. Credibility assessment of digital news sources is improved.

c. The mere awareness of potential digital deception constitutes part of new media literacy and can prevent undesirable consequences.

d. The proposed veracity/deception criterion is also seen as a metric for information quality assessment.

2.1.2 Disadvantages of Paper

a. This method works in the context of textual news information only.

2.1.3 How to overcome the problems mentioned in Paper

a. To target other than textual means enabling support for images as well, the images may more influence than text, as it may involve visual scenes which may be obscene in nature, or not usual for normal user, indirectly it can play with reader's intellect.

b. The images involved in the form of images can be scraped by using Google's Image Vision API, and then from the results we can extract origin, date of the image, as in fake news the image must have been used from past incident or any other incident must be framed as recent. Also, the description of searched results can be compared with the input news.

2.2 Automatic Deception Detection - Methods for Finding Fake News

News verification aims to employ technology to identify intentionally deceptive news content online, and is an important issue within certain streams of library and information science.

This paper provides researchers with a map of the current landscape of veracity deception assessment methods, their major classes and goals, all with the aim of proposing a hybrid approach to system design. These methods have emerged from separate development streams, utilizing disparate techniques. In this survey, two major categories of methods emerge:

1. Linguistic Approaches in which the content of deceptive messages is extracted and analyzed to associate language patterns with deception; and
2. Network Approaches in which network information, such as message meta-data or structured knowledge network queries can be harnessed to provide aggregate deception measures.

Both typically incorporate machine learning techniques for training classifiers to suit the analysis. It is incumbent upon researchers to understand these different areas, yet no known typology of methods exists in the current literature. The goal is to provide a survey of the existing research while proposing a hybrid approach, which utilizes the most effective deception detection methods for the implementation of a fake news detection tool.

Structured data sets are easier to verify than non-structured (or semi-structured) data such as texts. When we know the language domain (e.g., insurance claims or health-related news) we can make better guesses about the nature and use of deception. Semi-structured non-domain specific web data come in many formats and demand flexible methods for veracity verification.

2.2.1 Advantages of Paper

a. Success was measured based on whether the machine was able to assign higher true values to true statements than to false ones.
b. Linguistic and network-based approaches have shown high accuracy results in classification tasks within limited domains.

2.2.2 Disadvantages of Paper

a. A problem with this method is that statements must reside in a preexisting knowledge base.

2.2.3 How to overcome the problems mentioned in Paper

a. By using Hybrid approach that is combination of linguistic and network approach we can overcome the above mention problem.

2.3 Automatic Detection of Fake News

Fake news detection has recently attracted a growing interest from the general public and researchers as the circulation of misinformation online increases. For instance, a recent report by the Jumpshot Tech Blog found that Facebook referrals accounted for 50% of the total traffic to fake news sites and 20% total traffic to

reputable websites. Since the majority of U.S. adults –62%– gets news on social media, being able to identify fake content in online sources is a pressing need.

There are three important lines of research into the automated classification of genuine and fake news items. On a conceptual level, a distinction has been made between 'three types of fake news: serious fabrications (i.e. news items about false and nonexistent events or information such as celebrity gossip), hoaxes (i.e. providing false information via, for example, social media with the intention to be picked up by traditional news websites) and satire (i.e. humorous news items that mimic genuine news but contain irony and absurdity).

In this paper, we develop computational resources and models for the task of fake news detection. We present the construction of two novel datasets covering seven different domains. One of the datasets is collected using a combination of manual and crowdsourced annotation efforts, while the second is collected directly from the web. Using these datasets, we conduct several exploratory analyses to identify linguistic properties that are predominantly present in fake content, and we build fake news detectors relying on linguistic features that achieve accuracies of up to 78%. To place our results in perspective, we also compare the accuracy of our fake news detection models with an empirical human baseline accuracy.

2.3.1 Advantages of Paper

a. Datasets are collected manually and through crowdsourcing which means indirectly the chances of accuracy is high.

b. Introduced two new fake news datasets, one obtained through crowdsourcing and covering six news domains, and another one obtained from the web covering celebrities.

c. Most importantly at the end, fake news model are compared with the considerable empirical evidence to place our results in perspective.

2.3.2 Disadvantages of Paper

a. Distinction is done only between three types of news, nowadays there are many manipulative techniques to swing the news.

b. Collection of datasets is totally relied on crowdsourcing and web, which is not a good sign, stuffs obtained from these must be cross checked many a times.

c. As we can see the accuracy level is 78% which is OK OK result, but not the best, accuracy must be more 90% to get the tag of trustworthy application.

2.3.3 How to overcome the problems mentioned in Paper

a. We will have to keep a threshold value for calculating fake or genuine news. For example, if threshold value is 40, and above 40% users say that news is fake, then only it will be considered fake else it will be considered genuine.

2.4 Personalized News Filtering and Summarization on the Web

Information on Web pages is rapidly inflated and congested with large amounts of news contents. To identify useful information that satisfies a user's interests, the filtering and summarization of personalized Web news have drawn much attention in Web intelligence. The filtering and summarization of personalized Web news refer to the recommendation, extraction, and summarization of interesting and useful information from Web pages, which can be widely used to promote the automation degree in public opinion investigation, intelligence gathering and monitoring, topic tracking, and employment services.

This paper presents a personalized news filtering and summarization (PNFS) system that works on news Web pages. The first task of our system is to recommend interesting news to users. We dynamically obtain Web news from the Google news website (http://news.google.com), and then recommend personalized news to the users according to their interest preferences. A news filter is applied in our system to provide high quality news content for analyzing. The second research component of the PNFS system is to summarize Web news. The summarization is given in the form of keywords based on lexical chains.

A Web news recommendation mechanism is provided according to the users' interests which makes our PNFS system specially designed for personalized news treatment. An embedded learning component interacts with the recommendation mechanism and models user's interests. A keyword knowledge base is stored to update the user's profile, and a keyword extraction algorithm is also provided to construct the lexical chains based on word sense disambiguation.

1. **Recommender Systems:** There are mainly three different techniques commonly used in recommender systems: content-based recommendation, collaborative filtering, and hybrid recommendation.

   - The content-based approach recommends items based on the profile that is built by analyzing the content of articles that a user has read in the past.
   - The collaborative filtering approach uses the known preferences of a group of users to make recommendation for other users.
   - Hybrid approaches combine content-based methods with collaborative filtering techniques, aiming to avoid the limitations of each approach and improve the recommendation performance.

2. **Web News Extraction:** Web information extraction can be traced back to the integration research of heterogeneous data sources of structured and semi-structured data. A wrapper is viewed as a component in an information integration system to encapsulate accessing operations of multiple heterogeneous data sources, with which users can query on the integration system using a single uniform interface. As information extraction is the key function in a wrapper, the terms extractors and wrappers are often used interchangeably.

3. **Keyword Extraction:** Research in keyword extraction began in early 1950's. Existing work can be categorized into two major approaches: supervised extraction and unsupervised extraction. Supervised methods view keyword extraction as a classification task, where labeled keywords are used to learn a model. This model is constructed using a set of features that capture the saliency of a word as a keyword. Turney designed a keyword extraction system GenEX based on C4.5. Naive Bayes to extract keywords, and designed the Kea system.

2.4.1 Advantages of Paper

a. Recommended news is generated automatically when user reads a news on Google news based on some keyword from that google news which user has read.

b. Also in the recommended news, further filtering is done and new recommended news is generated on the previous recommended news. So continuously, filtering of recommended news is done as per user's preference.

2.4.2 Disadvantages of Paper

a. This paper gives recommendation only if user reads news on "Google news" and not on any other website.

b. Some keywords that are stored in keyword knowledge base is not removed if the user doesn't like it, due to which same recommended news will be shown every time.

2.4.3 How to overcome the problems mentioned in Paper

a. If the user is not reading some news based on some keyword, that keyword should be removed from keyword base after some period of time.

2.5 Intelligent News Aggregator and Validator

The news is a source which provides daily information about technology, sports, political, information etc. In recent years people most often interact with news using their smart phones, and there are many platforms like Facebook, Google, Twitter which easily provide news as per user interest.

In social media or on Internet not all news are genuine, sometimes the news propagated is made intentionally fake. Fake news, to be precise is a misleading information or a fabricated story about a public figure, political leader or a party or about certain issues. Users consume these fake news without properly assessing the content of the news and fall for the ulterior motive of fake news propagator. This sometimes affect the readers as well as society.

To mitigate the problem of fake news many fake news detection technique and methods were proposed in the past. Many of them works as well with certain percentage of accuracy. Reference [1] uses naive Bayes classifier for detection of fake news, and using this classification technique to detect fake news, the efficiency of 74% was obtained. Reference [4] uses two approach linguistic approach and network approach, in linguistic approach the contents of deceptive messages is extracted and analyzed to recognize the certain language patterns of such deceptiveness and in network approach, network information, such as network queries and structured knowledge are utilized to provide aggregate deception measures. Reference [5] With the help of the News data sets several analyses were conducted to identify linguistic properties that are present in fake news content and fake news detectors gaining accuracy up-to 70% were used. And at-last to be more perspective the accuracy of fake news detection models are compared with human baseline accuracy.

In our proposed system we will be addressing this fake news problem by providing users with a news platform/application from where they can get all their news, along with the validation of the news they are reading.

Our system will provide users every category of news ranging from politics to sports to pop culture to local news. These news will scraped from online news sources. The scraping of news data will be done periodically in small intervals in order to provide users with the latest news all the time. A learning algorithm will be used to provide users with personalized news. Personalized news will be based on user preferences and viewing history. The user can provide rating of news they are reading as genuine or fake. This rating can be used later on in the validation process.

2.5.1 Advantages of Paper

a. Users can access any type of news content without switching to another application.

b. Users get real-time genuine percentage of news articles.

c. The application dynamically validate news items as news content gets updated.

2.5.2 Disadvantages of Paper

a. As data set grow day by day, the validation mechanism can take little bit more time to validate news content.

b. As users can provide feedback on the respective news item. Classification of genuine users can become complex procedure.

2.5.3 How to overcome the problems mentioned in Paper

a. Only validated news content will be shown to users.

b. Make genuine user list, from which the genuine feedback is accepted for validation procedure.

c. Maintain good environment for application to work without any fluctuations.

2.6 Technical Review

The technology we're using here in implementation of our described project are: React js, PWA, MongoDb Atlas, Machine learning, Bag of Words model, News API, Heroku. All the listed technologies are described further.

2.6.1 React js

React is a front-end library developed by Facebook. It is used for handling the view layer for web and mobile apps. ReactJS allows us to create reusable UI components. It is currently one of the most popular JavaScript libraries and has a strong foundation and large community behind it.

React is a library for building composable user interfaces. It encourages the creation of reusable UI components, which present data that changes over time. Lots of people use React as the V in MVC. React abstracts away the DOM from you, offering a simpler programming model and better performance. React can also render on the server using Node, and it can power native apps using React Native. React implements one-way reactive data flow, which reduces the boilerplate and is easier to reason about than traditional data binding.

2.6.2 Reasons to use this Technology

a. Uses virtual DOM which is a JavaScript object. This will improve apps performance, since JavaScript virtual DOM is faster than the regular DOM.

b. Can be used on client and server side as well as with other frameworks.

c. Component and data patterns improve readability, which helps to maintain larger apps.

2.6.3 Progressive Web App

Progressive Web Apps (PWA) are experiences that combine the best of the web and the best of apps. They are useful to users from the very first visit in a browser tab, no install required. As the user progressively builds a relationship with the app over time, it becomes more and more powerful. It loads quickly, even on flaky networks, sends relevant push notifications, has an icon on the home screen, and loads as a top-level, full screen experience.

2.6.4 Reasons to use this Technology

a. Progressive Web Apps (PWA) match the mobile-first approach. This means that stores developed with PWA will function on mobile devices without any issues.

b. PWA combines the UX of a native app with the benefits of the mobile web. Stores with PWA run smoothly both as a web page and native app.

2.6.5 MongoDb Atlas

MongoDB Atlas is a fully-managed cloud database developed by the same people that build MongoDB. Atlas handles all the complexity of deploying, managing, and healing your deployments on the cloud service provider of your choice (AWS, Azure, and GCP). Follow the links below to get started.

MongoDB Atlas provides the functionality and reliability you need, with its built-in automation mechanisms. With MongoDB Atlas, you no longer need to worry about operational tasks.

2.6.6 Reasons to use this Technology

a. Automation at the Core

b. Flexibility & Support

c. Security

d. Scalability

e. Highly Available

f. High Performance

2.6.7 Machine learning

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that which makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

2.6.8 Reasons to use this Technology

a. Easily identifies trends and patterns

b. No human intervention needed (automation)

c. Continuous Improvement

d. Handling multi-dimensional and multi-variety data

e. Wide Applications

2.6.9 Bag of Words model

Bag of words model is one of a series of techniques from a field of computer science known as Natural Language Processing or NLP to extract features from text. The way it does this is by counting the frequency of words in a document. A document can be defined as you need, it can be a single sentence or all Wikipedia. The output of the bag of words model is a frequency vector.

In simple terms, it's a collection of words to represent a sentence with word count and mostly disregarding the order in which they appear. BOW is an approach widely used with:

- Natural language processing
- Information retrieval from documents
- Document classifications

On a high level, it involves the following steps:

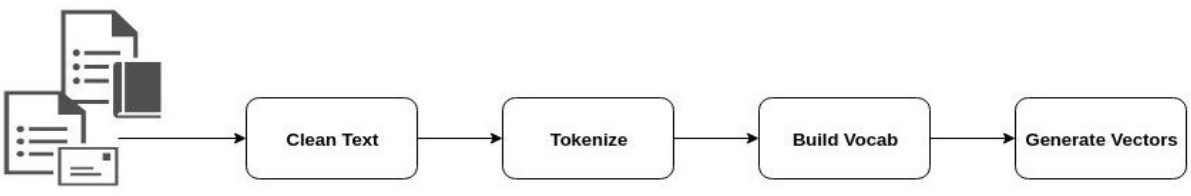


Figure 2.1: Bag of words model

2.6.10 Reasons to use this Technology

a. Its very efficient technique.

b. Its more responsive text classification method. All the processes are executed sequentially.

2.6.11 News API

News API is a simple HTTP REST API for searching and retrieving live news articles from all over the web. Using this, one can fetch the top stories running on any news website or can search top news on specific topic (or keyword).

News can be retrieved based on some criteria. Say the topic (keyword) to be searched is 'Geeksforgeeks' or might be concerned to specific channel. All can be done, but API Key is needed to get started.

2.6.12 Reasons to use this Technology

a. It returns JSON metadata we can be easily use.

b. It response time is good.

c. Huge number of data sources that is over 30,000 news sources.

2.6.13 Heroku

Heroku is a cloud platform as a service (PaaS) supporting several programming languages. Heroku, one of the first cloud platforms, has been in development since June 2007, when it supported only the Ruby programming language, but now supports Java, Node.js, Scala, Clojure, Python, PHP, and Go. For this reason, Heroku is said to be a polyglot platform as it has features for a developer to build, run and scale applications in a similar manner across most languages.

2.6.14 Reasons to use this Technology

a. Heroku is easy to get started, you only need to install heroku ruby gem.

b. Heroku is the cheapest option for a low traffic site.

c. They offer no of Dyno for upgrade and downgrade app instance.

d. Database integration is pretty simple with PostgreSQL.

Chapter 3

Project Planning

3.1 Members and Capabilities

Table 3.1: Table of Capabilities

SR. No	Name of Member	Capabilities
1	Khan Shahid Jahid	Validation & Rest Server
2	Prasad Chandresh Kamleshwar	Database Design
3	Khatri Amaanulla Abdul Sattar	Testing
4	Khan Irfan Abubakar	UI Design & Rest Server

3.2 Roles and Responsibilities

Table 3.2: Table of Responsibilities

SR. No	Name of Member	Role	Responsibilities
1	Khan Shahid Jahid	Team leader	Core Modules
2	Prasad Chandresh Kamleshwar	Team Member	Database Design
3	Khatri Amaanulla Abdul Sattar	Team Member	Testing
4	Khan Irfan Abubakar	Team Member	UI Design

3.3 Assumptions and Constraints

3.3.1 Assumption:

Assumption is that the data that is coming from the scrapper is in correct from and from genuine source.

3.3.2 Constraint:

If news API from where we are scrapping the data goes down then the system will fail.

3.4 Project Management Approach

In **"The Waterfall"** approach, the whole process of software development is divided into separate phases. The outcome of one phase acts as the input for the next phase sequentially. This means that any phase in the development process begins only if the previous phase is complete. The waterfall model is a sequential design process in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of Conception, Initiation, Analysis, Design, Construction, Testing, Production/Implementation and Maintenance.

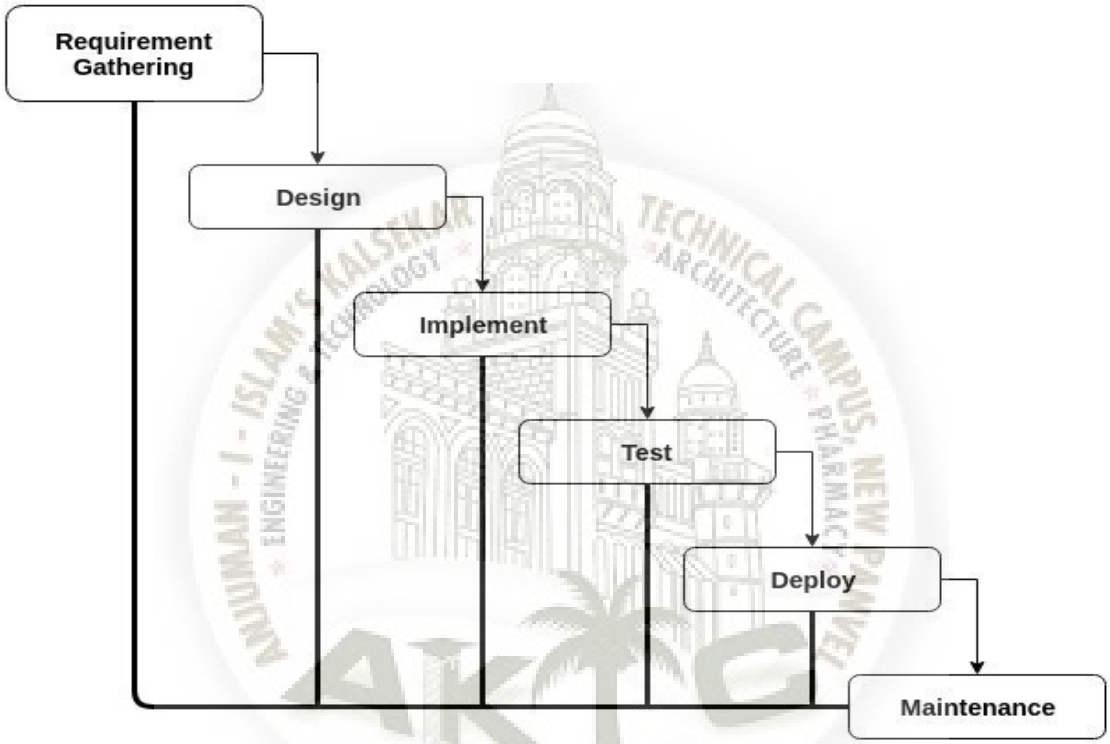


Figure 3.1: Waterfall Model

3.5 Ground Rules for the Project

1. We treat each other with respect.
2. We intend to develop personal relationships to enhance trust and open communication.
3. We value constructive feedback. We will avoid being defensive and give feedback in a constructive manner.
4. As team members, we will pitch in to help where necessary to help solve problems and catch-up on behind schedule work.

5. Additional meetings can be scheduled to discuss critical issues or tabled items upon discussion and agreement with the team leader.
6. One person talks at a time, there are no side discussions.
7. When we pose an issue or a problem, we will also try to present a solution.

3.6 Project Budget

The budget for this project is very low as most of the tools we have use are open source. Most of the tools are freely available across the internet.

3.7 Project Timeline

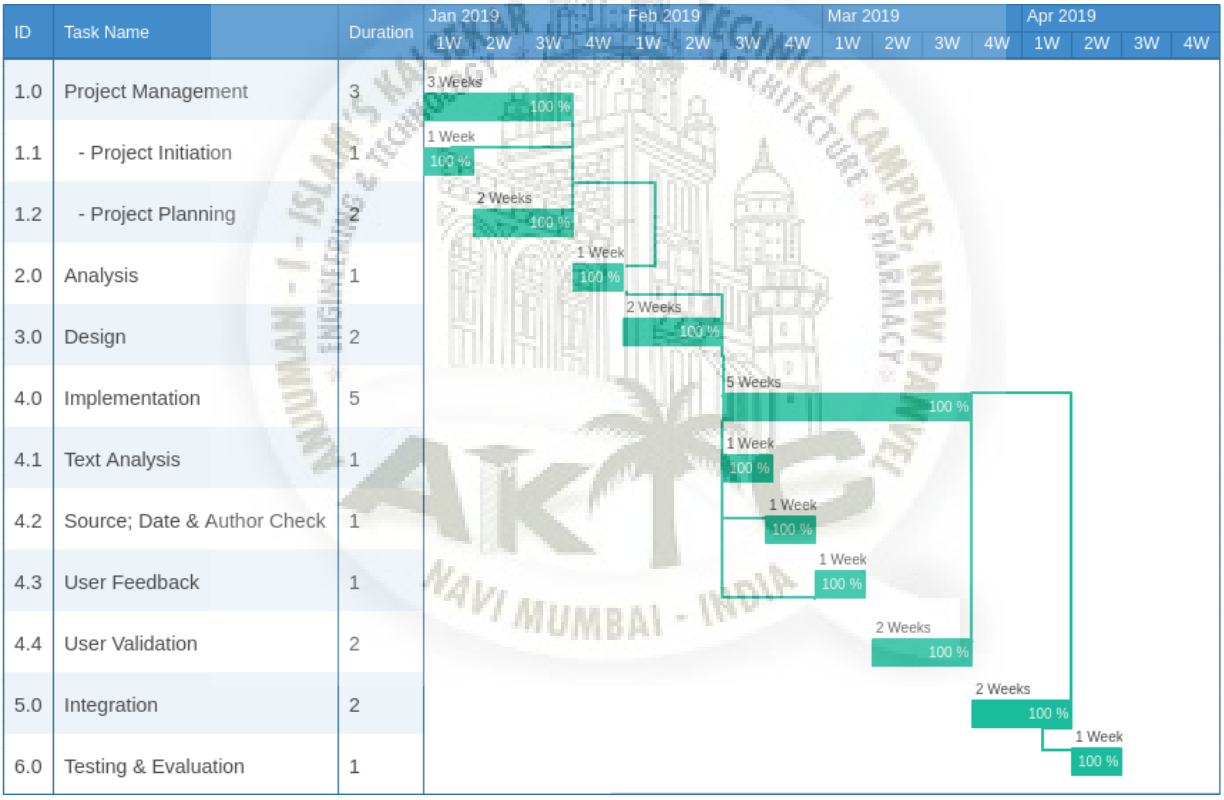


Figure 3.2: Gantt Chart

Chapter 4

Software Requirements Specification

4.1 Overall Description

4.1.1 Product Perspective

With the increasing amount of exposed information on the Internet, upon which everybody can post something, the evaluation of credibility and trustworthiness becomes difficult. The current phenomenon fake news does not make the access to reliable information easier. Because of the vast amount of information in the online world, which could make the time to evaluate every item limited, it is questioned whether this has an influence on credibility evaluations.

4.1.2 Product Features

This study aimed to examine this scientific gap by investigating the dependence of type of information processing on credibility evaluations of fake news. The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. The proliferation of misleading information in everyday access media outlets such as social media feeds, news blogs, and online newspapers have made it challenging to identify trustworthy news sources,thus increasing the need for computational tools able to provide insights into the reliability of online content.

For News Aggregation keyword knowledge base is maintained and provides a real-time update to reflect the general Web news topic information and the user interest preferences. The non-news content irrelevant to the news Web page is filtered out. Keywords that capture the main topic of the news are extracted using lexical chains to represent semantic relations between words. For fake news detection we are using Naive Bayes classifier (machine based validation)and User feedback (user based validation).

4.1.3 User Classes and Characteristics

There are various users that can use the our system such as normal user i.e any day to day life user who come across various news sources daily to get more knowledge

about whats going on in today's world. The content of the news is provided by our system is most likely accurate with its validation percentage. By which the user who accessing news content get brief idea about the news which he/she is following.

4.1.4 Operating Environment

Our system will be accessible through both WebApp as well as Mobile Application. Our system will be easy to understand and use.

4.1.5 Design and Implementation Constraints

The main decision was to select that which language should we use for developing our system. Since most of the programming language are very time consuming when it comes to data processing . So the complete system is made by using python language which is very efficient for processing the data as compare to the other language which are available in market since its provide various library to do our work simple.

4.2 System Features

In this project we are dealing with such type of circulated fake news on social sites and news portals. Our aim is straight forward, just classify and show each news as per their validation value. We are using recent technology such as machine learning as well as artificial intelligence to categories each news as per their validity.This is very much helpful for users to understand and make decision to follow which type of news articles for their reference. We are building this project to deal with fake news which is serious concern in this social generation to protect them from having misinformation about recent happenings.

4.2.1 Scrapping

The news data scraped from the internet will also serve another purpose in our system i.e. to be used by machine learning algorithms to be validated as fake or genuine. This validation information will later be used along with user rating to classify the news with even greater precision.

Description and Priority

It's is key player in our system it will extract news from over 4000 news website and get its data in java script object notation format to us.It is high priority module. Its main benefits is that it scrap the data from large number of reliable websites.

4.2.2 Machine Based Validation

For machine learning algorithm to predict/classify accurately the features used should produce meaningful results. For a news article there are quite a few traits that can be used as features.Some of those are author analysis, domain analysis, checking the content of the News text. As we can there are a lot features which on its own, independently can classify whether the News is fake or not to certain degree as seen in (1). A better approach is to consider all those features in a way that makes the classification easier. For all News articles we must generate parameters for each corresponding features. And use these generated parameters for classification. This will allow us to include as many meaning full features as possible to make classification even more accurate. The generated parameters generated will be scaled between 0 and 1. Number of features used will be explained in detail in later section.

4.2.3 Classification Done Twice

As mentioned before the goal of our project is to build a validation system that is dynamic. For the validation system to be dynamic two basic requirements should be fulfilled that is the size of the training data set should be increasing and the entropy in the data should also increase.This will allow our system to predict the validity of the newer News with greater accuracy.To achieve the objective of inflating the training set, what we can do is add the newly predicted News into the training set. But this creates a problem, if the data that is added to increase the size of training set is predicted from the training set itself. The new data that is being added is not distinguishable from the original training set. The entropy in the data does not increase when newer News data is added, hence failing the second requirement. For increasing the training set, adding the newly predicted data is fine as long as there is another way of verifying whether the prediction made by machine learning algorithms is correct. That's where the user rating of News comes in. It provided provides another alternative for verifying whether the prediction made the machine learning algorithm is right or wrong. Our system will compare the News rating provided by users and News validation calculated by machine learning, and save the result of comparison in the training set. But taking user rating of News for validation comes with certain problem as well. In the next section we will discuss those problem and how we try to overcome it.

4.2.4 Verification Of User Ratings

When users provide the rating for the News we have to make sure that, before using it for validation we filter out all the ratings that are inaccurate, biased, purposefully wrong, or provided by the users to mislead the system. For that purpose our system will filter out user who are providing inaccurate ratings. This will be

done in two steps; first step will be ranking users based on their accuracy ratio and second step will find out patterns in the wrong output of top ranked users, if certain pattern exists then ignore the rating from such user. More details of these two steps are explained in the later sections. This two process will help us fetch out News rating which are cream of the crop, and with least possible errors.

Stimulus/Response Sequences

1. The User must register in our system.
2. After Registration user must login in our system using its login credentials.In response to login the system will provide a API key.
3. By using API key provided by the system user can input for the data to validate.
4. In response to enter data the system will predict whether the enter data is correct or false. And it will return a value.
5. In case if the data is correct it will display that which is the source of the data.

Functional Requirements

1. The user should register on the system User should login in the system.
2. The data input by the user must be in textual from.
3. The servers should response quickly.

4.3 External Interface Requirements

4.3.1 User Interfaces

1. User must first register itself on our system.
2. After registration user will get a API key through will he/she will be authenticated.
3. By using that key user can search for the data he/she needs to validate.

4.3.2 Hardware Interfaces

There is no special kind of hardware required to run this system, A regular desktop or any smart phone can be enough to access our system.

4.3.3 Software Interfaces

One must need a web browser or our mobile application to access our system. For database we used No sql database that is MongoDB. For developing core modules we used python programming language. We have used news API for extracting the required data. On the client side only a browser or our mobile application can be used.

4.3.4 Communications Interfaces

1. The major communication between the client or user with our system is done using web-browser or an mobile application provided by us.
2. The communication between the system and database is done by using MongoDB.

4.4 Nonfunctional Requirements

4.4.1 Performance Requirements

The main performance of this system depend on how much time it spends on extracting and processing the data. extracting the data from various news sources is very time consuming task but python made it simple by providing library. And also for processing the used language that is python is very efficient.

4.4.2 Safety Requirements

Once the client login in our system he/she can provide his/her valuable feedback to the news items which get published on our app. Any random user cannot provide the feedback for misguiding our machine based system.

4.4.3 Security Requirements

For security purpose all the internal operations will be monitored by authorised person. Only genuine users are provided with feedback facility. And if any authorised user who tries to fool our system by giving random feedback (like for genuine news, 'fake' or fake news as 'genuine') will be blocked as per rules defined.

Chapter 5

System Design

5.1 System Requirements Definition

System requirement definitions specify what the system should do, its functionality and its essential and desirable system properties. The techniques applied to elicit and collect information in order to create system specifications and requirement definitions involve consultations, interviews, requirements workshop with customers and end users. The objective of the requirements definition phase is to derive the two types of requirement: The objective of the requirements definition phase is to derive the two types of requirement:

5.1.1 Functional requirements

They define the basic functions that the system must provide and focus on the needs and goals of the end users.

Use-case Diagram

Use case diagrams are usually referred to as behavior diagrams used to describe a set of actions (use cases) that some system or systems (subject) should or can perform in collaboration with one or more external users of the system (actors). Each use case should provide some observable and valuable result to the actors or other stakeholders of the system.

Use case diagrams are in fact two fold they are both behavior diagrams, because they describe behavior of the system, and they are also structure diagrams as a special case of class diagrams where classifiers are restricted to be either actors or use cases related to each other with associations.

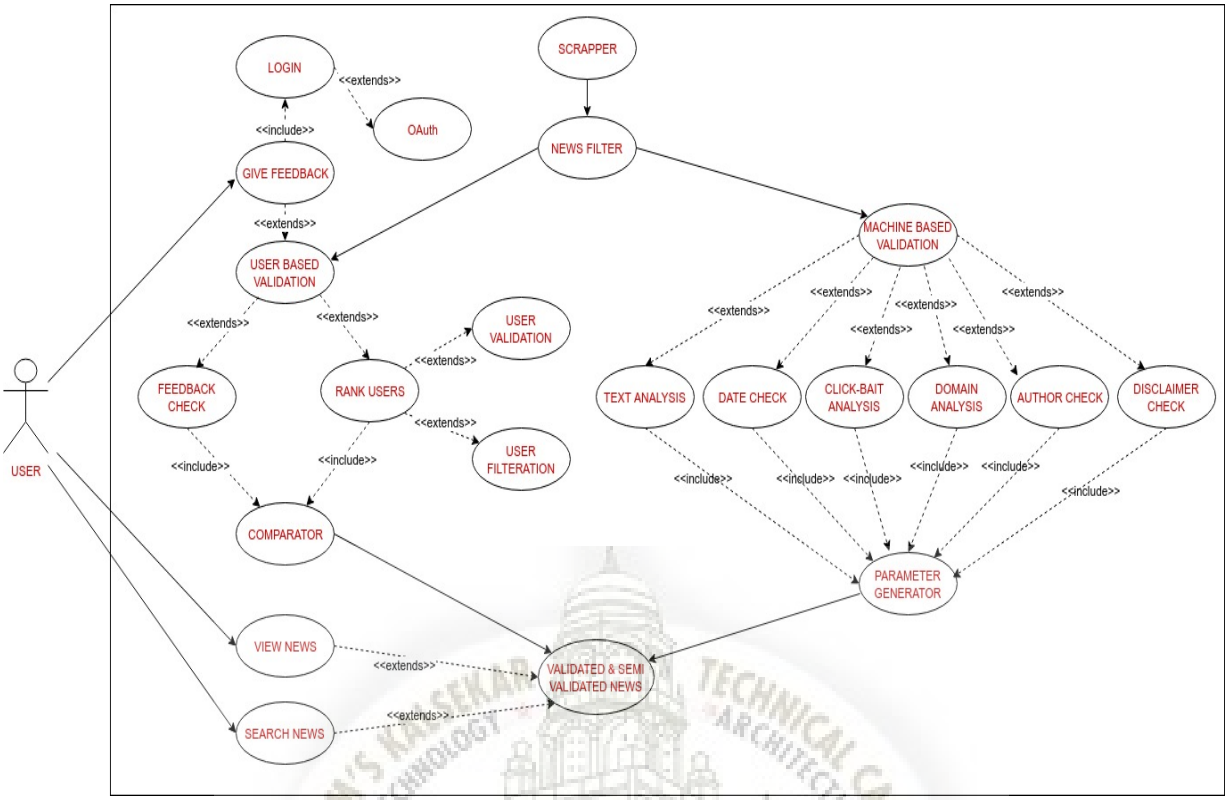


Figure 5.1: Use Case Diagram

Data-flow Diagram

A data flow diagram (DFD) maps out the flow of information for any process or system. It uses defined symbols like rectangles, circles and arrows, plus short text labels, to show data inputs, outputs, storage points and the routes between each destination. Data flowcharts can range from simple, even hand-drawn process overviews, to in-depth, multi-level DFD's that dig progressively deeper into how the data is handled. They can be used to analyze an existing system or model a new one. Like all the best diagrams and charts, a DFD can often visually "say" things that would be hard to explain in words, and they work for both technical and nontechnical audiences, from developer to CEO. That's why DFD's remain so popular after all these years. While they work well for data flow software and systems, they are less applicable nowadays to visualizing interactive, real-time or database-oriented software or systems.

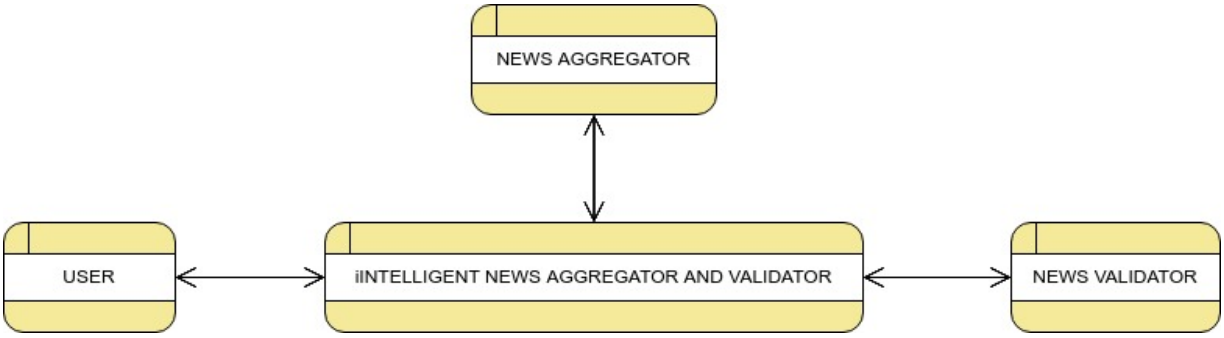


Figure 5.2: Level 0 Data Flow Diagram

1.jpg

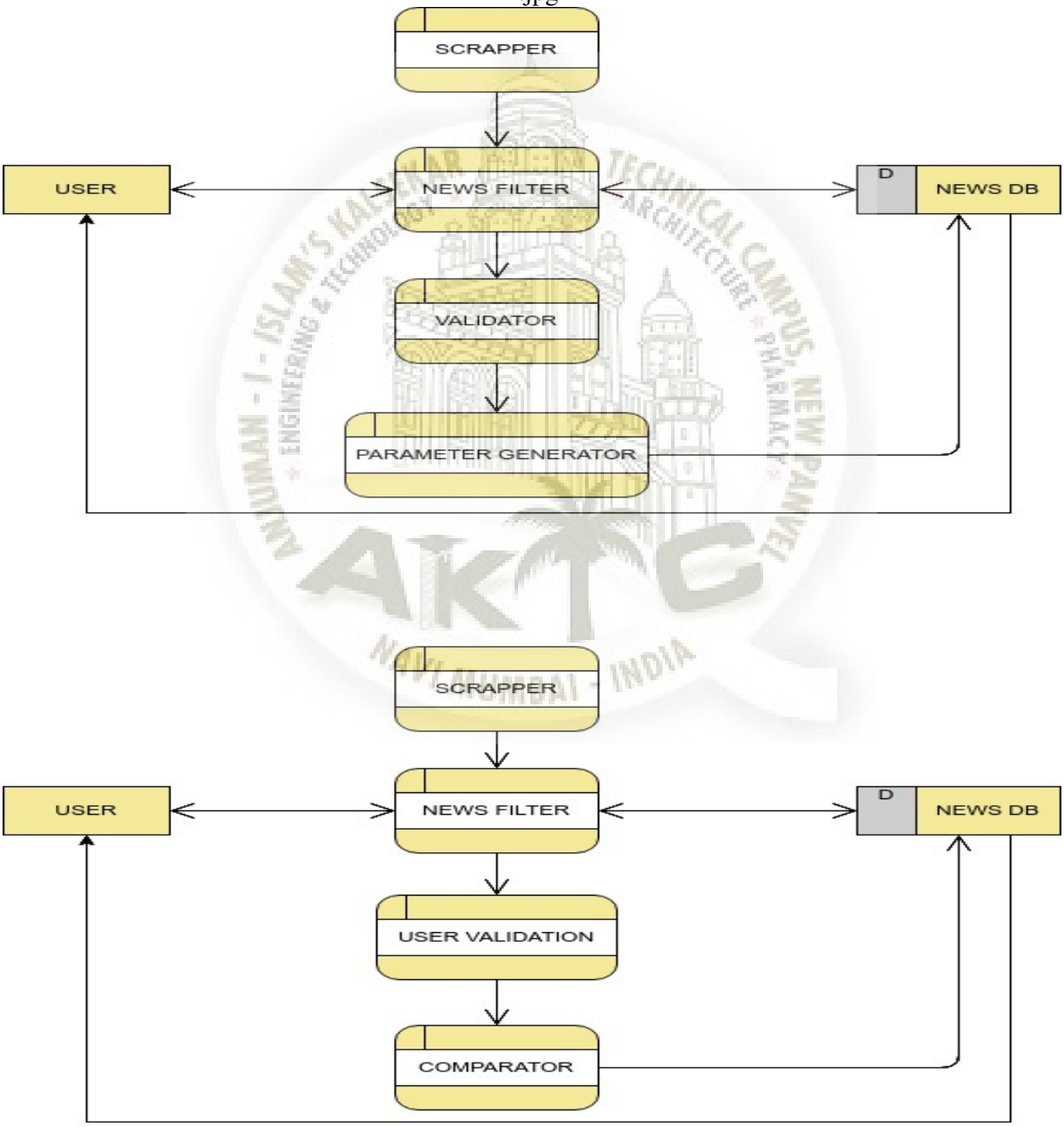


Figure 5.3: Level 1 Data Flow Diagram

2.jpg

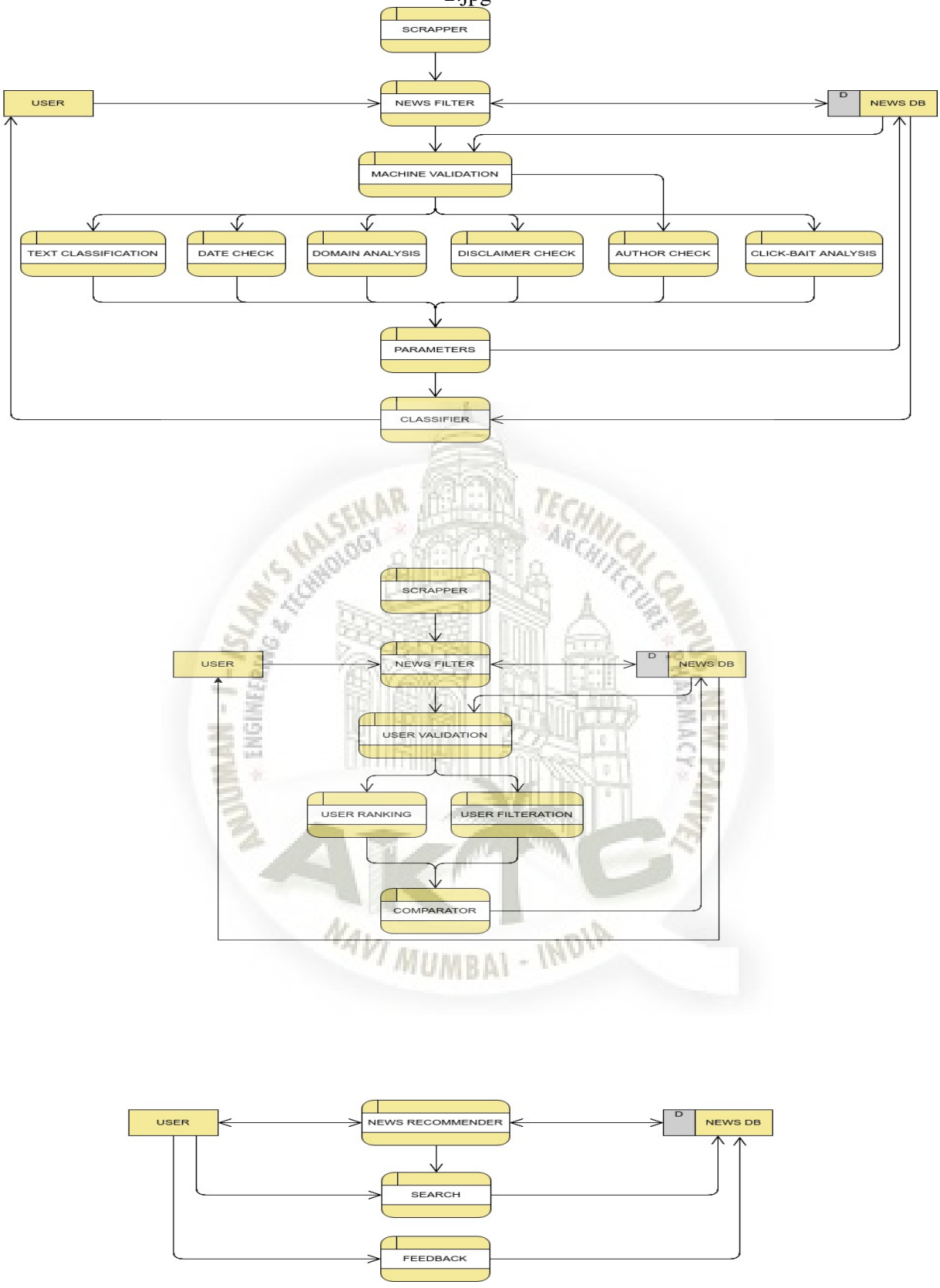


Figure 5.4: Level 2 Data Flow Diagram

5.1.2 System requirements (non-functional requirements)

Performance Requirements

The main performance of this system depend on how much time it spends on extracting and processing the data. extracting the data from various news sources is very time consuming task but python made it simple by providing library. And also for processing the used language that is python is very efficient.

Safety Requirements

Once the client login in our system he/she can provide his/her valuable feedback to the news items which get published on our app. Any random user cannot provide the feedback for misguiding our machine based system.

Security Requirements

For security purpose all the internal operations will be monitored by authorised person. Only genuine users are provided with feedback facility. And if any authorised user who tries to fool our system by giving random feedback (like for genuine news, 'fake' or fake news as 'genuine') will be blocked as per rules defined.

Database Schema/ E-R Diagram

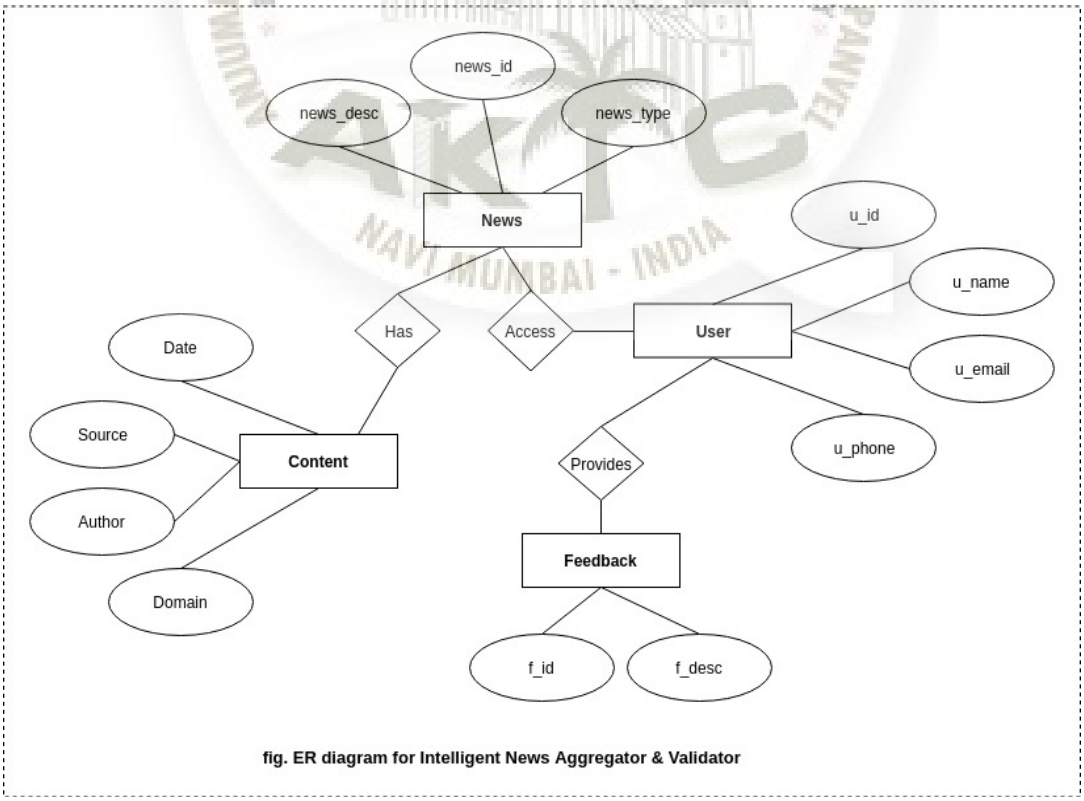


Figure 5.5: Entity Relationship Diagram

5.2 System Architecture Design

As stated in the introduction the goal of our system is to provide users a platform where they can easily access the news they want to read along the validity of the news. In order to achieve this goal certain operations are needed to be performed. These operations are done in done in six phases, these six phases are explained in depth in later section but the list of those operation are as follows: Scrapping news from Internet and filtering out important data from the news page, Extracting keywords from news titles, Generation of parameters for classification, Classification using machine learning, Providing users personalized News, Getting News rating from users for User based Classification, Comparison and classification using user rating and machine learning based classification.

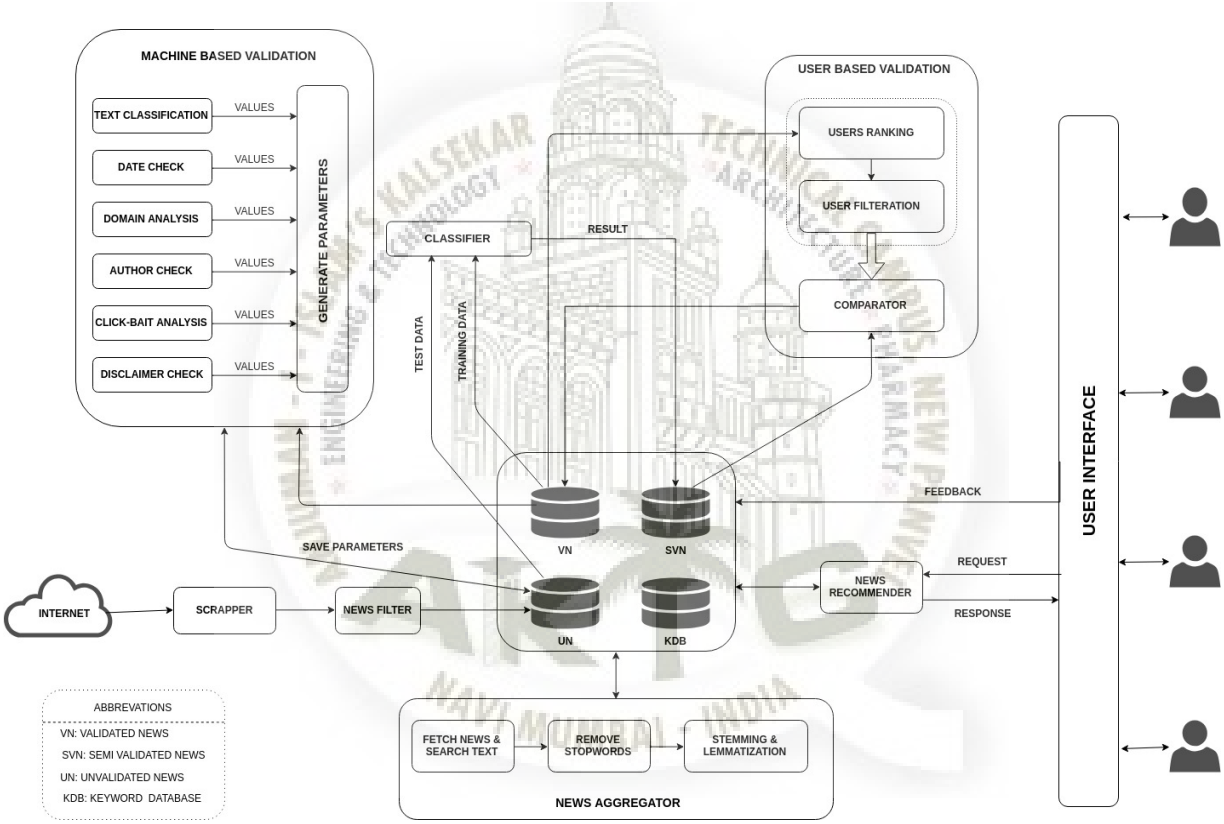


Figure 5.6: System Architecture

5.3 Sub-system Development

There are total six main modules in system architecture namely Scrapper, News Aggregator, Machine Based Validation, Classifier, News Recommendation Component, User Based Validation. All the modules will be briefly described further:

5.3.1 Scrapper

- **Scrapper:** This component helps fetch News from various sources, from the internet.
- **News Filter:** With the help of this component, we are going to discard all the news that are already stored in database, and save the News which are not present in the database.

5.3.2 News Aggregator

- **Fetch News and Search Text:** In this component we fetch news titles and the saved searches which were made by users from the database of whom the keywords were not extracted.
- **Remove Stop words:** In this component we remove the stop words from the fetched News titles so that only the useful words remains. For Example: consider the list of string ['i', 'am', 'going', 'to', 'go', 'to', 'the', 'store', 'and', 'park'] after removing the stop words, the list of words will only contain ['going', 'go', 'store', 'park'].
- **Stemming and Lemmatization:** This Component will reduce the words to root words from derived and complex words. For Example: Stemming handles matching "car" to "cars". Lemmatization handles matching "car" to "cars" along with matching "car" to "automobile".

5.3.3 Machine Based Validation

Once the unvalidated news is fetched from the internet, our system will generate and save parameters of those news and save in database for later use in classifier.
Parameter Values: The parameter values as mentioned before will be between (0 and 1).

- **Text Classification:** This Component uses the bag-of-words model and Naive Bayes classifier to check how the content(text) of the news article relate to other true or fake news articles.
- **Click bait Analysis:** This component checks how much the news title is related to the news article using recurrent neural networks.
- **Domain Analysis:** This component checks if domain from which the news belong is good or not. This will be done by getting the accuracy ratio of that domain from the database.
- **Date check:** This component checks whether the news article is an old news article that is being republished.

- **Author Check:** This component checks the accuracy ratio of past articles written by the author.
- **Disclaimer Check:** The component check at the bottom of the article if it makes a disclaimer about accuracy not being good.

5.3.4 Classifier

This Component uses saved parameters of validated news set for training the classification model. And use the saved parameter of unvalidated news for testing/predicting the validity of the news and save in semi-validated News set.

5.3.5 News Recommendation Component

Fetches keywords from database and user's browsing history to recommend news according to the users liking and preference. Only News from validated news and semi-validated news data set will be served. Machine learning Algorithms such as Naive Bayes classifier and K Nearest Neighbour will used for this component.

5.3.6 User Based Validation

We provide the users News from validated and semi-validated news data set for reading and ask them what is their feedback on the news that they are reading. This will give user's rating on validated news which will help us filter out inaccurate, biased, purposefully wrong feedback providing users out. And use News rating of users that were not filtered out after ranking and pattern recognition.

- **User Ranking:** This component ranks users based on their accuracy ratio. The accuracy ratio is obtained by taking the ratio of their correct News rating versus incorrect News rating. The question arrives here how do we figure out, whether the News rating provided by the user is correct or not. This can be obtained by comparing user's News rating with News in Validated News set. If the News rating of the user matches with the Validity of the validated News set then that rating is correct and vice versa. After this fetch News Rating of the users that comes in top 5-10% by accuracy ratio.
- **User filtration by Pattern:** The users obtained after ranking and elimination; we do a pattern recognition on the titles of wrong News rating of those users to see if some pattern emerges, if it is does then eliminate those user as well. After this ranking and filtration we can expect to have feedback with least possible inaccuracy.
- **Comparator:** Compare the prediction obtained from the News rating of the users and semi-validated news, and save the news which comes as true and

false similarly in both cases in Validated News set. And in cases in which cases there is contradiction start the process again from phase 3. (goal for doing this is to see if new validated data may improve/change the machine prediction or user rating may change with time).

5.4 Systems Integration

There are mainly six modules in our system, in order to integrate complete system certain operations are needed to be performed. These operations are done in done in six phases, these six phases are explained in depth in later section but the list of those operation are as follows: Scrapping news from Internet and filtering out important data from the news page, Extracting keywords from news titles, Generation of parameters for classification, Classification using machine learning, Providing users personalized News, Getting News rating from users for User based Classification, Comparison and classification using user rating and machine learning based classification.

5.4.1 Class Diagram

This is the Class diagram of the system in which the modules which will be there after the deployment are shown. This class diagram is an illustration of the relationships and source code dependencies among classes in the Unified Modeling Language (UML). In this context, a class defines the methods and variables in an object, which is a specific entity in a program or the unit of code representing that entity.

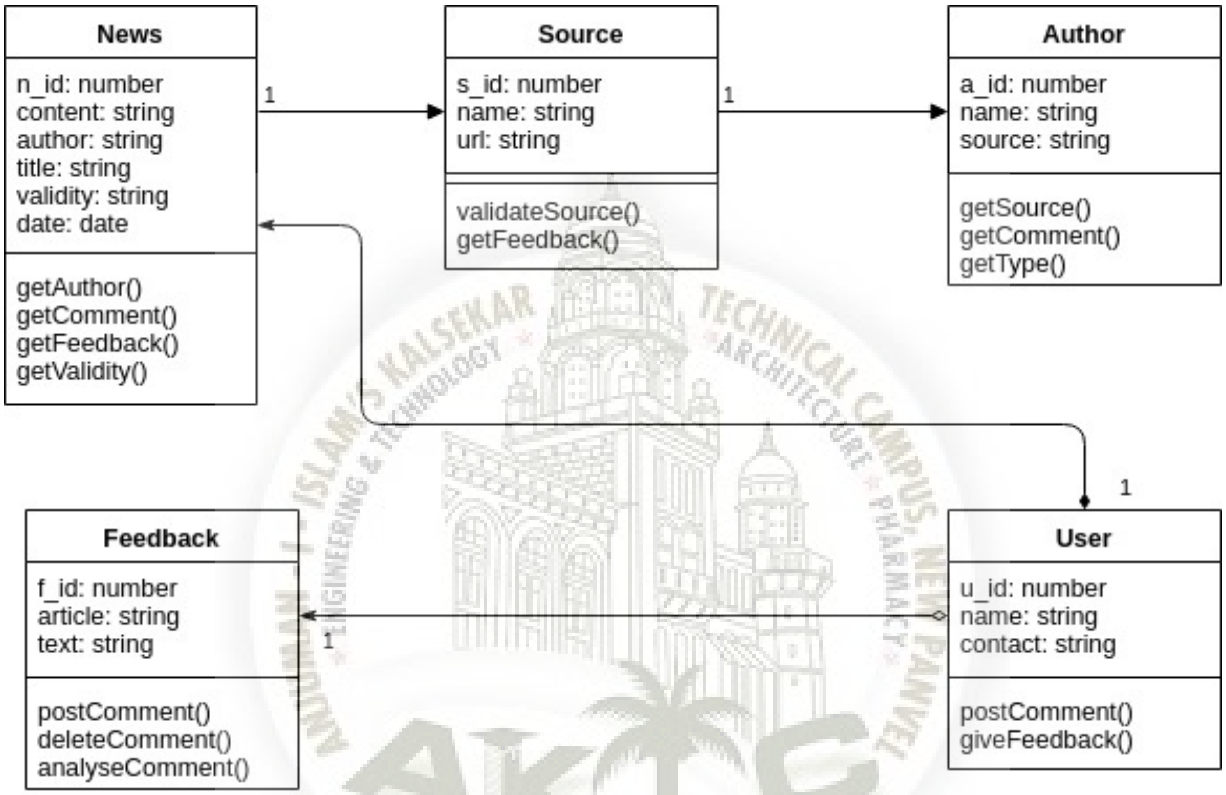


Figure 5.7: Class Diagram

5.4.2 Sequence Diagram

To understand what a sequence diagram is, it's important to know the role of UML. UML, or the Unified Modeling Language, is a modeling toolkit that guides the creation and notation of many types of diagrams, including behavior diagrams, interaction diagrams, and structure diagrams. Sequence diagrams are a kind of interaction diagram, because they describe how and in what order a group of objects works together. These diagrams are used by software developers and business people alike to understand requirements for a new system or to document an existing process. Sequence diagrams are sometimes known as event diagrams or event scenarios.

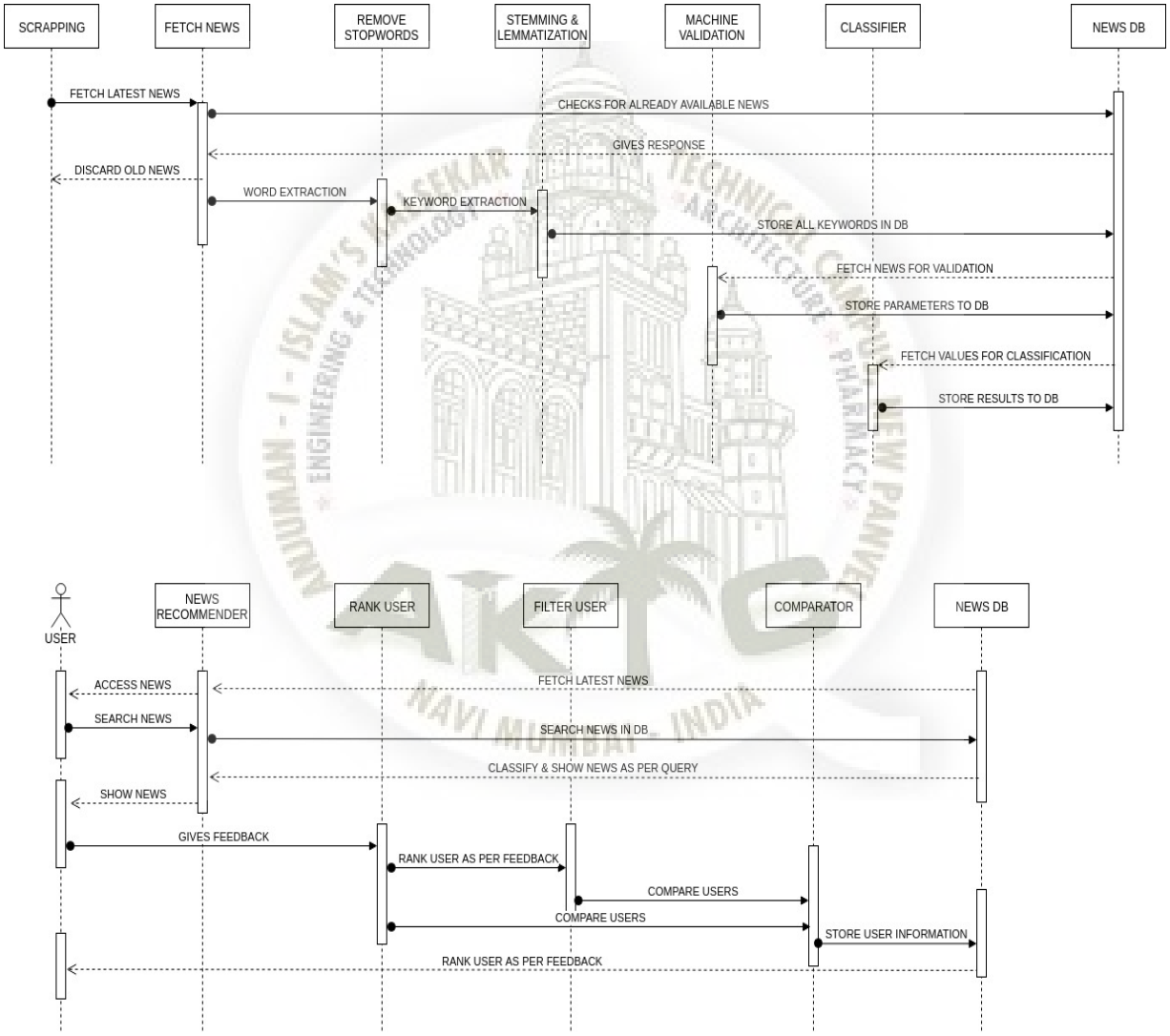


Figure 5.8: Sequence Diagram

5.4.3 Component Diagram

Component diagrams are different in terms of nature and behavior. Component diagrams are used to model the physical aspects of a system. Now the question is, what are these physical aspects? Physical aspects are the elements such as executables, libraries, files, documents, etc. which reside in a node.Component diagrams are used to visualize the organization and relationships among components in a system. These diagrams are also used to make executable systems.

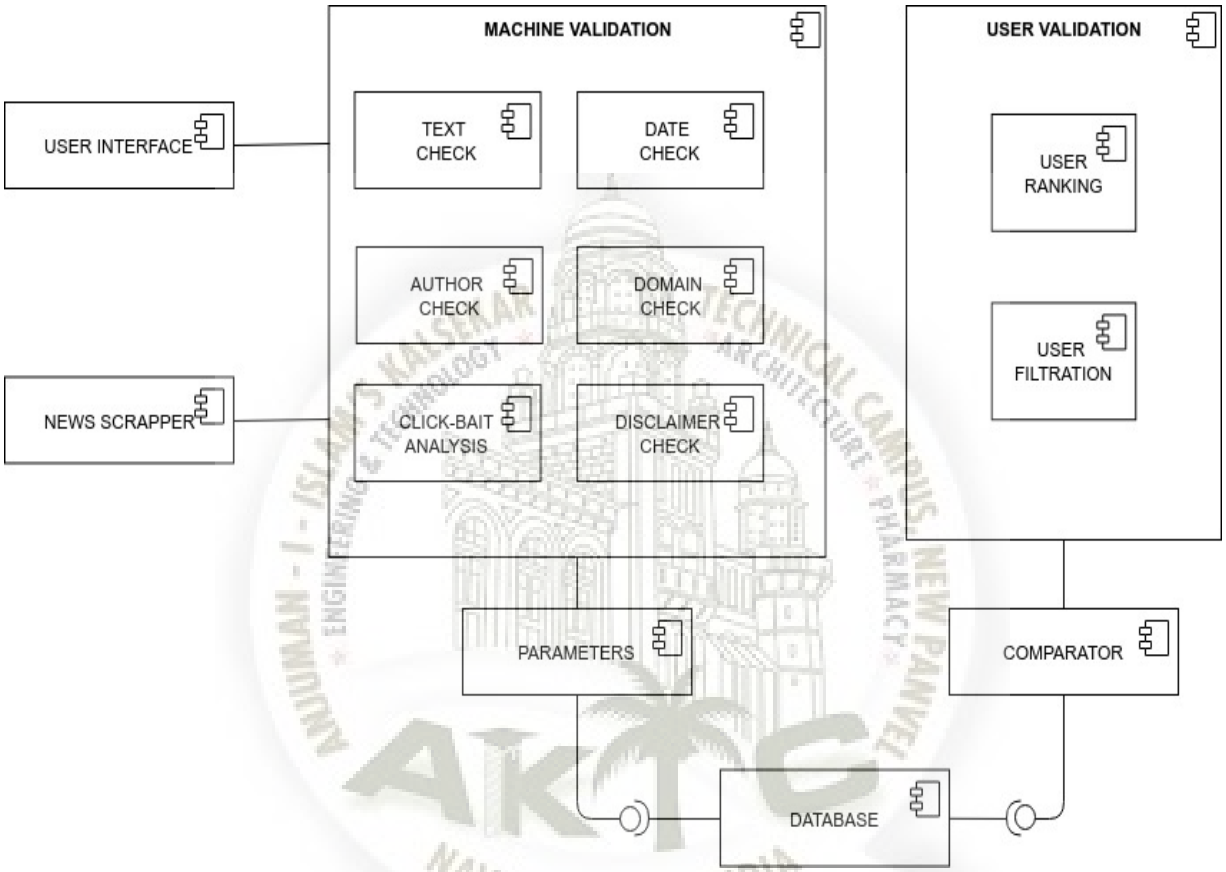


Figure 5.9: Component Diagram

5.4.4 Deployment Diagram

Deployment diagram is a structure diagram which shows architecture of the system as deployment (distribution) of software artifacts to deployment targets. The below figure describe the Deployment Diagram of our system.

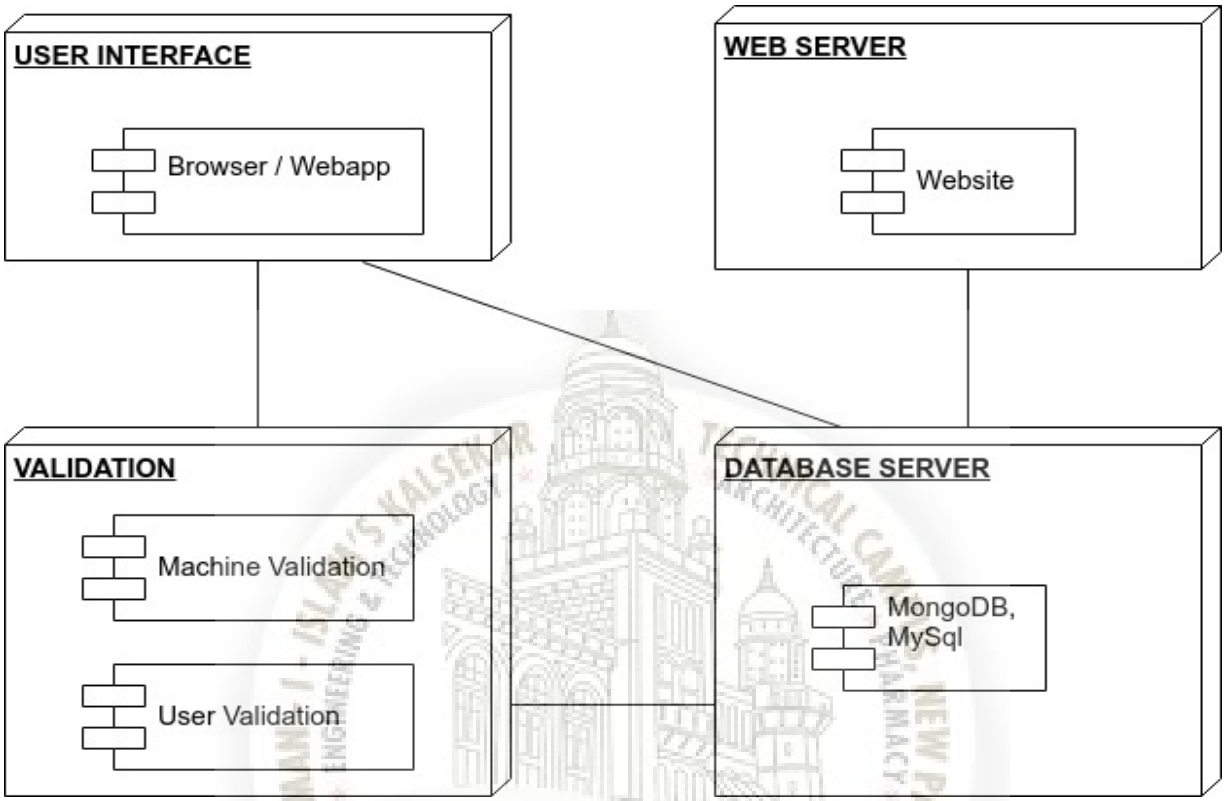


Figure 5.10: Deployment Diagram

Chapter 6

Implementation

6.1 Scrapper Module

The news data scraped from the internet will also serve another purpose in our system i.e. to be used by machine learning algorithms to be validated as fake or genuine. This validation information will later be used along with user rating to classify the news with even greater precision.

```
class Fetch(object):

    @classmethod
    def connect(cls, key):
        cls.handle = NewsApiClient(api_key=key)

    @classmethod
    def get_top_news(cls, cat):
        articles = cls.handle.get_top_headlines(country='in', category=cat,
                                                page_size=100)
        return articles['articles']


class Store(object):

    @classmethod
    def check_already_exist(cls, title):
        news = News.objects.raw({'title': title})
        if news.count() > 0:
            return False
        return True

    @classmethod
    def save_news(cls, titl, auth, desc, urll, iurl, pubd, sorc, cont, cate):
        news = News(title=titl, url=urll, img_url=iurl, author_name=auth,
                    source_name=sorc, published_at=pubd, content=cont,
                    description=desc, category=cate)
        news.save()
        return (news._id)

    @classmethod
    def add_to_categories(cls, newslist, cat):
        try:
            c = Categories.objects.get({"cat_name": cat})

```

```
            except DoesNotExist:
                Categories(cat_name=cat).save()
                c = Categories.objects.get({"cat_name": cat})

            c.news_list.extend(newslist)
            c.save()

    @classmethod
    def add_to_invalidated_news(cls, ids):
        try:
            newslist = HelperLists.objects.first()
        except DoesNotExist:
            newslist = HelperLists().save()
            newslist = HelperLists.objects.first()
        newslist.unvalidated_news.extend(ids)
        newslist.save()
```

6.2 News Aggregator Module

In this module we fetch news titles and the saved searches which were made by users from the database of whom the keywords were not extracted.

Next we remove the stop words from the fetched News titles so that only the useful words remains. For Example: consider the list of string ['i', 'am', 'going', 'to', 'go', 'to', 'the', 'store', 'and','park'] after removing the stop words, the list of words will only contain ['going', 'go', 'store', 'park'].

Than will reduce the words to root words from derived and complex words. For Example: Stemming handles matching "car" to "cars". Lemmatization handles matching "car" to "cars" along with matching "car" to "automobile".

```
connect('mongodb://localhost:27017/news2')
# connect(f'mongodb+srv://{Secret.mongo_user}:{Secret.mongo_password}@'
# 'newsbu-a20e5.mongodb.net/news?retryWrites=true')


class News(MongoModel):
    title = fields.CharField(required=True)
    url = fields.URLField()
    img_url = fields.URLField()
    author_name = fields.CharField()
    source_name = fields.CharField()
    published_at = fields.DateTimeField()
    content = fields.CharField()
    description = fields.CharField()
    category = fields.CharField()
    validity = fields.IntegerField()


class Keyword(MongoModel):
    newsid = fields.ReferenceField(News)
    title = fields.CharField(blank=True)
    content = fields.CharField(blank=True)
    author = fields.CharField(blank=True)
    source = fields.CharField(blank=True)


class Categories(MongoModel):
    cat_name = fields.CharField()
    news_list = fields.ListField(fields.ReferenceField(News))
```

6.3 Machine Based Validation Module

For machine learning algorithm to predict/classify accurately the features used should produce meaningful results. For a news article there are quite a few traits that can be used as features.Some of those are author analysis, domain analysis, checking the content of the News text. As we can there are a lot features which on its own,independently can classify whether the News is fake or not to certain degree as seen. A better approach is to consider all those features in a way that makes the classification easier. For all News articles we must generate parameters for each corresponding features. And use these generated parameters for classification. This will allow us to include as many meaning full features as possible to make classification even more accurate. The generated parameters generated will be scaled between 0 and 1.

```
class TC(object):  # Text classification

  class Actions(object):
      classifier = GaussianNB()
      cv = CountVectorizer()
      X = []
      x_train = []
      x_test = []
      y_train = []
      y_pred = []

      @classmethod
      def vectorize(cls, corpus):
          cls.X = cls.cv.fit_transform(corpus).toarray()

      @classmethod
      def prep_train_test(cls, test):
          cls.x_test = cls.X[-test:]
          cls.x_train = cls.X[:-test]

      @classmethod
      def prep_validity(cls, news):
          cls.y_train
          for i in news:
              cls.y_train.append(i.validity)

      @classmethod
      def train_model(cls):
          cls.classifier.fit(cls.x_train, cls.y_train)

      @classmethod
      def predict(cls):
          cls.y_pred = cls.classifier.predict(cls.x_test)
          return cls.y_pred
```

6.4 Classifier Module

This Component uses saved parameters of validated news set for training the classification model. And use the saved parameter of unvalidated news for testing/predicting the validity of the news and save in semi-validated News set.

```
class Actions(object):
    classifier = GaussianNB()
    cv = CountVectorizer()
    X = []
    x_train = []
    x_test = []
    y_train = []
    y_pred = []



    @classmethod
    def prep_train_test(cls, test):
        cls.x_test = cls.X[-test:]
        cls.x_train = cls.X[:-test]

    @classmethod
    def prep_validity(cls, news):
        cls.y_train
        for i in news:
            cls.y_train.append(i.validity)

    @classmethod
    def train_model(cls):
        cls.classifier.fit(cls.x_train, cls.y_train)

    @classmethod
    def predict(cls):
        cls.y_pred = cls.classifier.predict(cls.x_test)
        return cls.y_pred

    @classmethod
    def calculate(cls, news):
        a,s,t = (news.author_check,
                 news.source_check,
                 news.text_classification)
        return (a+s+t)/3
```

6.5 User Based Validation Module

When users provide the rating for the News we have to make sure that, before using it for validation we filter out all the ratings that are inaccurate, biased, purposefully wrong, or provided by the users to mislead the system. For that purpose our system will filter out user who are providing inaccurate ratings. This will be done in two steps; first step will be ranking users based on their accuracy ratio and second step will find out patterns in the wrong output of top ranked users, if certain pattern exists then ignore the rating from such user. More details of these two steps are explained in the later sections. This two process will help us fetch out News rating which are cream of the crop, and with least possible errors.

```
class FeedbackValidator{

    constructor(){
        this.userid
        this.all_user_feedbacks
        this.validated_news_list
        this.user_validated_news_feedback
        this.newsValidities = []
        this.userValidities = []
        this.accuracy = 0
        this.ratio = 0
    }

    // fetch the all the feedback for that user
    async get_user_feedbacks(){
        try{
            this.all_user_feedbacks = await Feedback.find({
                "userid": this.userid
            })
            // console.log(this.all_user_feedbacks)
        } catch (err) {
            console.log(err)
        }
    }

    // weed out news that are not validated
    async is_valid_news(){
        try{
            let lists = await Lists.find({})
            this.validated_news_list = lists[0].validated_news
        }catch (err) {
            console.log(err)
        }
        let i = 0
        let newsid
        let newsList = []
        let validity
        while(i<this.all_user_feedbacks.length){
            newsid = this.all_user_feedbacks[i].newsid
            validity = this.all_user_feedbacks[i].validity
            if (this.validated_news_list.indexOf(newsid) > -1){
                newsList.push(newsid)
                this.userValidities.push(validity)
            }
```

```
            i++
        }
        this.user_validated_news_feedback = newsList
    }

    async fetch_news_validities(){
        let i = 0
        let validity
        while(i<this.user_validated_news_feedback.length){
            validity = await News.findById(
                this.user_validated_news_feedback[i]
            )
            this.newsValidities.push(validity.validity)
            i++
        }
    }

    // compare users feedback with all the validated news
    async compare_validities(){
        let i = 0
        while(i<this.userValidities.length){
            if (this.userValidities[i] == this.newsValidities[i]){
                this.accuracy++
            }
            i++
       }
    }
    async calculate_accuracy(){
        this.ratio = this.accuracy / this.user_validated_news_feedback.length
    }
    async save_ratio_to_db(){
        let user = await User.findById(this.userid)
        user.accuracy = this.ratio
        user.save()
    }

    flush_variables(){
        this.userid = 0
        this.all_user_feedbacks = []
        this.validated_news_list = []
        this.user_validated_news_feedback = []
        this.newsValidities = []
        this.userValidities = []
        this.accuracy = 0
        this.ratio = 0
    }

    async run(userid){
        this.userid = userid
        await this.get_user_feedbacks()
        await this.is_valid_news()
        await this.fetch_news_validities()
        await this.compare_validities()
        await this.calculate_accuracy()
        await this.save_ratio_to_db()
        this.flush_variables()
    }

}
```

6.6 Front end

```
const NewsItem = (props) => {
  const { img_url, title, source_name, description,published_at,validity,url } =
      props;

  let isValid= ""
  if (validity === undefined) {
    isValid = <span className="badge badge-warning">Not Validated </span>
  } else {
    isValid = validity === 0 ?
              <span className="badge badge-danger">Likely to be Fake</span> :
              <span className="badge badge-success">Likely to be Genuine</span>
  }

  return (
    <Row className="my-1 mx-1 py-3 bg-white newsCol">
      <Col xs='8'>
        <Link to={`/news/${props._id}`} className="text-dark text-decoration-
            none">
          <p className="h5 font-weight-bold mb-0 font-smaller">{title}</p>
        </Link>
        {isValid}
        <small className="text-muted d-block text-truncate mb-0" style={{width:"
            80%"}}>{description}</small>
        <div className="d-flex justify-content-between">
          <div className="d-flex flex-column">
            <small className="font-weight-bold mb-0">
              <a href={url} className="text-dark">
                {source_name}
              </a>
            </small>

            <small className="text-monospace text-muted">
              <Moment format="MMM DD,YYYY">
                {published_at}
              </Moment>
            </small>
          </div>
        </div>
      </Col>
      <Col xs='4' className="d-flex justify-content-sm-end">
      <Link to={`/news/${props._id}`} >
        <img src={img_url}
             style={{width:"100%",maxWidth:'150px', objectFit: 'cover',height: '
                 100px'}}
             onError={(e)=>{e.target.onerror = null; e.target.parentNode.
                 parentNode.parentNode.firstChild.style="max-width:100% !
                 important;flex:none;"; e.target.parentNode.parentNode.style="
                 display:none"}}
             alt={title}
        />
      </Link>
      </Col>
    </Row>
  );
};
```

```
class NewsDetail extends React.Component {
  state = {
    tooltipOpen: false,
    feedbackShow:false
  }

  componentDidMount(){
    if(this.props.match.params.id){
      this.props.fetchNewsDetail(this.props.match.params.id)
    }
  }


  toggle=()=>{
    this.setState({
      tooltipOpen: !this.state.tooltipOpen
    });
  }

  togglefeedback=async ()=>{
    await this.setState({
      feedbackShow: !this.state.feedbackShow
    });
    if(this.state.feedbackShow){
      this.props.fetchNewsFeedback(this.props.match.params.id)
    }
  }

  componentWillUnmount(){
    this.props.clearFeedback();
  }

  render(){
    const { newsDetail, loading, } = this.props;
    const {isAuthenticated}= this.props.auth;
    const {newsFeedback , loading:feedbackLoading} =this.props.feedback;
    console.log(this.props.feedback);


    let feedbackAuthCheck ='';
    if(isAuthenticated){
      if(feedbackLoading){
        feedbackAuthCheck='Loading...';
      }else if(newsFeedback){
        feedbackAuthCheck=<CardText>You rated this news to be {newsFeedback.
            validity?<Badge color="success">Genuine</Badge>: <Badge color="
            danger">Fake</Badge>}</CardText>
      }else{
        feedbackAuthCheck = (
          <>
          <CardText>How would you rate this news?</CardText>
          <div className="d-flex">
            <Button onClick={()=>this.props.addNewsFeedback(this.props.match.
                params.id,1)} color="success">Correct</Button>  
            <Button onClick={()=>this.props.addNewsFeedback(this.props.match.
                params.id,0)} color="danger">Fake</Button>
          </div>
          </>
          )
      }
    }else{
```

```
    feedbackAuthCheck = <CardText>To give feedback Please Login First. <Link
        className="auth-btn badge" to="/login"> Login</Link> </CardText>
  }

  // console.log(newsDetail,loading);
  const { img_url, title, source_name, description,published_at,validity,url }
      = newsDetail;

  let isValid= ""
  if (validity === undefined) {
    isValid = <span className="badge badge-warning">Not Validated </span>
  } else {
    isValid = validity === 0 ?
              <span className="badge badge-danger">Likely to be Fake</span> :
              <span className="badge badge-success">Likely to be Genuine</span
                  >
  }



  return (
    <MainContent className="news-post">
      <Container className="newsCol bg-white">
        <Row>
          <Col className="ml-0 mr-0 pr-0 pl-0">
            <img src={img_url}
              alt={title}
              onError={(e)=>{e.target.onerror = null;e.target.parentNode.
                  parentNode.style="display:none"}}
              className="mb-3 news-detail-img"
            />
          </Col>
        </Row>
        <Row>
          <Col className="mt-3 mb-3" sm="12" md={{ size: 8, offset: 2 }}>
            <h2 className="mb-2">{title}</h2>
            {isValid}
            {/* <p className="text-muted d-block mb-3">
            {description}</p> */}

            <p className="font-weight-bold mb-2"><a href={url} className="text
                -dark">
            {source_name}</a></p>

            {/* feedback dot btn */}
            <div className="float-right d-flex justify-content-center px-3 py
                -2" id="feedback" onClick={this.togglefeedback}>
              <i className="fa fa-ellipsis-v" aria-hidden="true"></i>
              <Tooltip placement="left" isOpen={this.state.tooltipOpen}
                  autohide={false} target="feedback" toggle={this.toggle}>
                Give your feedback
              </Tooltip>
            </div>

            <p className="text-monospace text-muted mb-3">
              <Moment format="MMM DD,YYYY">
                {published_at}
              </Moment>
            </p>
            <div className={classnames({"feedback-slideup":!this.state.
                feedbackShow},{"feedback-slidedown":this.state.feedbackShow})
```

```
                }>
163               <Card body>
164                 <CardTitle>Give Feedback</CardTitle>
165                   {feedbackAuthCheck}
166               </Card>
167             </div>
168
169           </Col>
170         </Row>
171
172         <Row className="mb-3 pb-3">
173           <Col sm="12" md={{ size: 8, offset: 2 }}  className="mb-3">
174           <p className="text-dark d-block mb-3">
175             {description}
176           </p>
177           <div>
178             <a href={url}>
179               Read more...
180             </a>
181
182           </div>
183
184           </Col>
185         </Row>
186       </Container>
187     </MainContent>
188   )
189  }
190 }
191
192 const mapStateToProps = state =>({
193     newsDetail : state.newsDetail.newsDetail,
194     loading :  state.newsDetail.loading,
195     auth:state.auth,
196     feedback:state.feedback,
197 })
198
199 const mapDispachToProps ={
200   fetchNewsDetail,
201   fetchNewsFeedback,
202   addNewsFeedback,
203   clearFeedback
```

Chapter 7

System Testing

As our proposed system is ready after implementation we test our application as per different system testing criteria.

7.1 Test Cases and Test Results

Test ID	Test Case Title	Test Condition	System Behavior	Expected Result
T01	Login	Should be registered user	User should login to our system	User presented with fresh Home-screen
T02	Search News	News must be in correct form	Gather news related to search details	Display the news content on screen
T03	Search non-existing news	News must be in correct form	Output should be NULL	Display ”No Results found” screen
T04	Provide feedback on news	Should be registered user	Calculate probability of feedback being genuine	Display user feedback with user rating

7.2 Sample of a Test Case

Title: Login Page – Authenticate successfully on our system.
Description: A registered user should be able to successfully login at our system.

Precondition: The user must already be registered with an email address and password.

Assumption: A supported browser/smartphone is being used.

Test Steps:

1. Navigate to Login window.

2. In the ’email’ field, enter the email and in ’password’ field enter Password of the registered user.
3. Click the ‘Login’ button.
4. If not registered user. Click ’Register’.

Expected Result: User successfully login to our system.

Actual Result: After successful login, User presented with home screen with fresh latest news articles across globe.

Figure 7.1: Application Homescreen

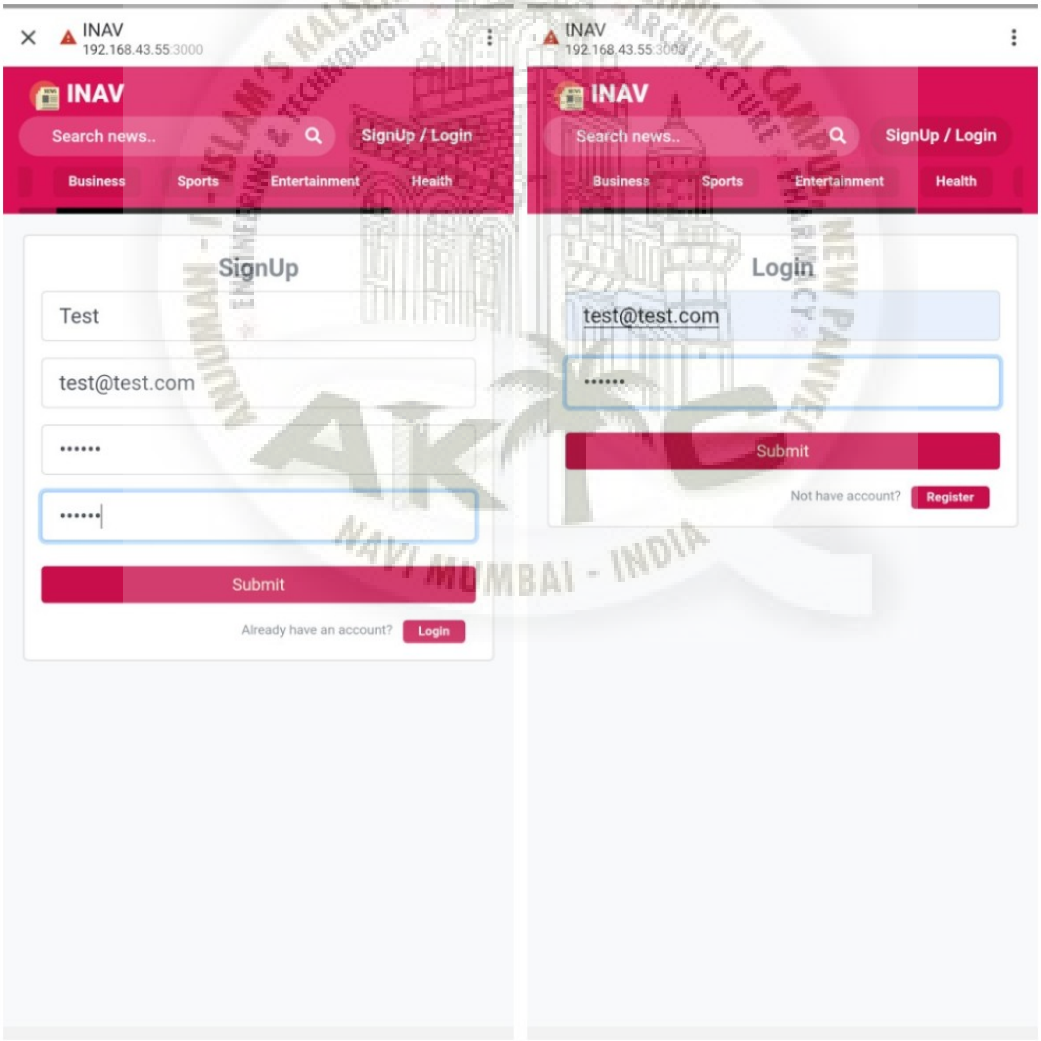
7.2.1 Software Quality Attributes

1. **Availability:** The system should not be down,whenever the user use the system the specific data should be available to the user.
2. **Correctness:** As per as the user search correct data should be shown to user.
3. **Maintainability:** The administrator of the system should maintain the system.
4. **Reliability:** The system should be reliable for producing correct output so that user can reliable on system.
5. **Extensibility:** The system is capable to be modified by changing some modules or by adding some features to the existing system.

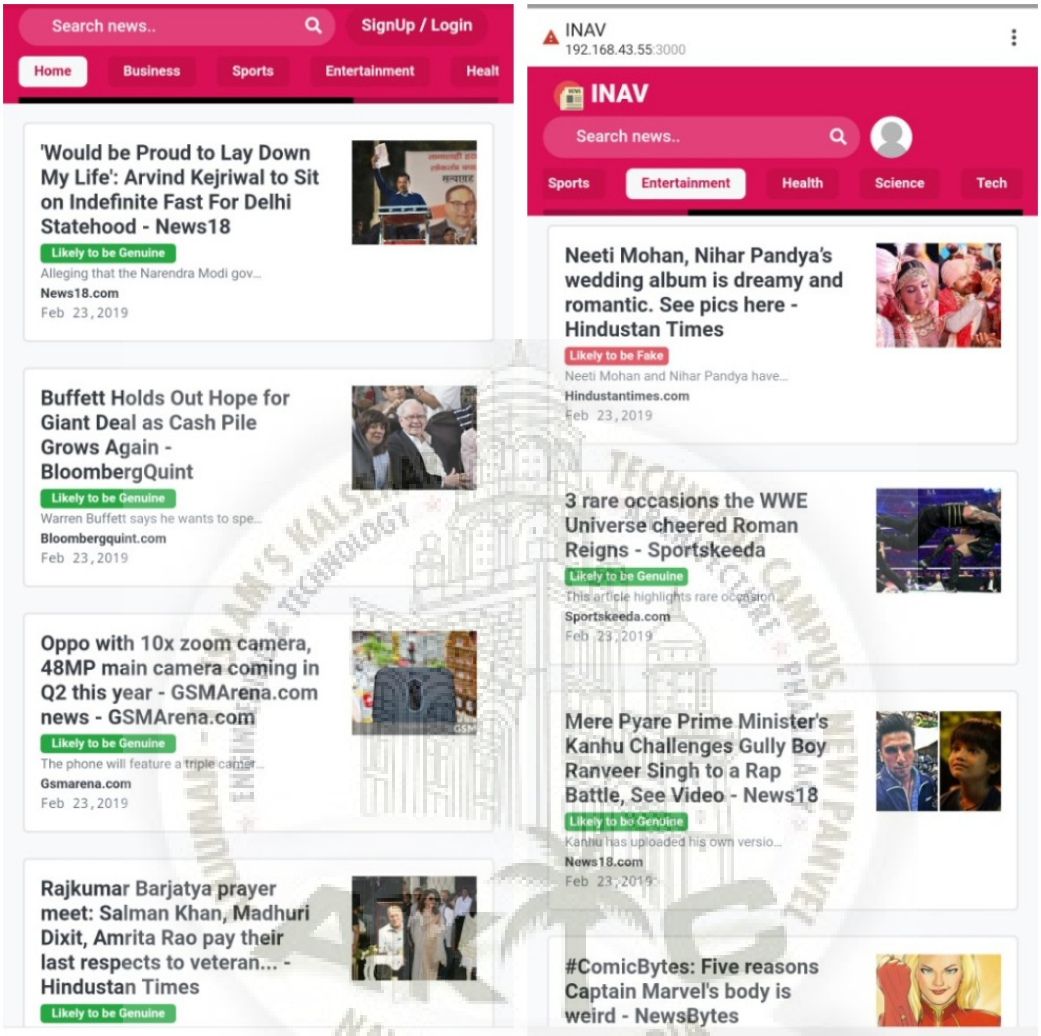
Chapter 8

Screenshots of Project

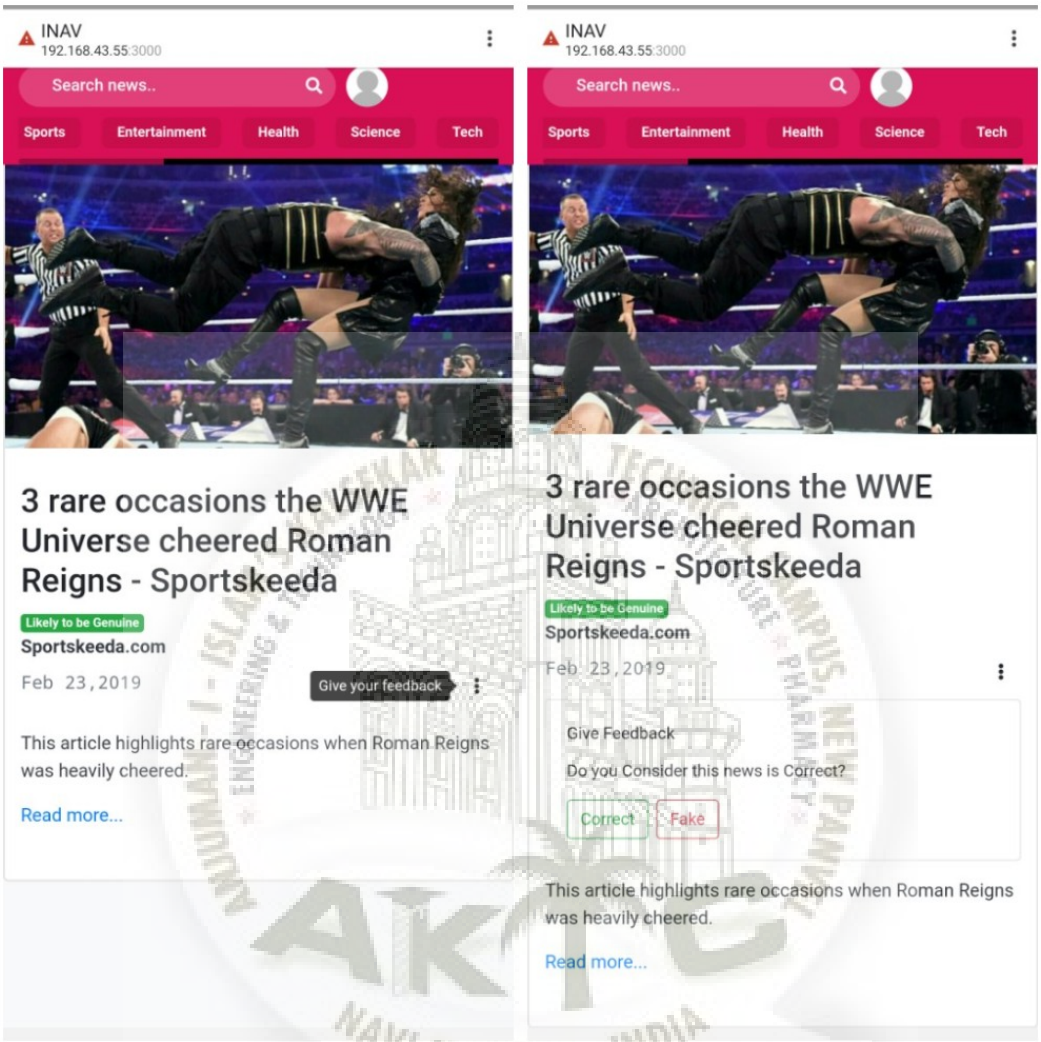
8.1 Login & Register



8.2 News Homescreen



8.3 News Description with feedback input



Chapter 9

Conclusion and Future Scope

9.1 Conclusion

In this project we inscribed the intelligent news aggregation and validation to provide and to distinguish between fake and genuine News, we have used machine learning and user rating for the overall completion of our news validation system. Our project whose implementation is done our application will be very helpful for any individual, with the help of our system users will be able to read the news and also will be able to check the validity of the news. We also calculated how accurate our validation system is and what flaws and error we are dealing with so that we can sort it out and make it flawless to gain absolute solution.

9.2 Future Scope

- In future, we plan to implement multiple accurate ways to validate news as well as users.
- We also planned to expand it to identify and validate news based on image, audio and video clips.

We are also planning to add some extra features or methods to generate more accurate results as our system grows. Some of the extra work we can do is:

9.2.1 Zipf's Law

Ranking of words is done depending of how frequently the appear in the text. In our project fake news text will be similarly ranked. Then any new fake new that arrives will be compared to these ranks.

9.2.2 Classification using neural networks

A custom classifier can be made using neural networks and our generated parameters can be passed into that neural network classifier.

9.2.3 Multimedia Analysis

Analysis of multimedia such as images, videos and audio can be done to verify it fake or not, or any old files is being republished by the publisher.

References

[1] Mykhailo Granik, Volodymyr Mesyur. Fake News Using Naive Bayes Classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering.

[2] Peter Adelson, Sho Aroro and Jeff Hara. Clickbait; Didn't read: Clickbait Detection using Parallel Neural Networks.

[3] Ankesh Anand, Tanmoy Charoborty, Noseong Park. We used Neural Networks to Detect Clickbaits: You Won't believe what happened Next?

[4] Niall J conroyy, Victoria L Rubin, Yimin Chen. Automatic Deception Detection: Methods for finding Fake News.

[5] Veronica Perez Rosas, Bennett Kleinberg, AlexandrLefevre, Rada Mihalcca. Automatic Detection of Fake News.

[6] Xindong Wu, Fei Xe, Gongqing Wu, Wei Ding. Personalized News Filtering and Summarization on the Web.

[7] Stemming https://en.wikipedia.org/wiki/Stemming

[8] Lemmatisation. https://en.wikipedia.org/wiki/Lemmatisation

[9] Machine Learing. https://en.wikipedia.org/wiki/Machine_learning

[10] Sufiyan Pawaskar, Nooralam Shaikh, Pradnyesh Rane and Tabrez Khan News Validation System.

Achievements

1. Publications

   (a) *Intelligent News Aggregator and Validator*; Shahid Khan, Chandresh Prasad, Amaanulla Khatri, Irfan Khan and Tabrez Khan 3rd Biennial International Conference on Nascent Technologies in Engineering (ICNTE 2019)

2. Conferences

   (a) *Intelligent News Aggregator and Validator*; Shahid Khan, Chandresh Prasad, Amaanulla Khatri, Irfan Khan and Tabrez Khan 3rd Biennial International Conference on Nascent Technologies in Engineering (ICNTE 2019) (Agnel Charities Fr. C. Rodrigues Institute of Technology)

Agnel Charities' Fr. C. Rodrigues Institute of Technology
Vashi, Navi Mumbai 400 703, India

3rd Biennial International Conference on Nascent Technologies in Engineering (ICNTE2019)

Certificate

This is to certify that Dr/Mr/Ms. Shahid Khan

has presented paper titled. Intelligent News Aggregator and Validator

in the ICNTE2019 organized by Fr.C. Rodrigues Institute of Technology, Vashi in its premises in association with IEEE on January 4-5, 2019.

IEEE
Advancing Technology for Humanity

ICNTE
Connecting ideas.

Dr. Nitesh [illegible]
Conference Chair

Dr. S.M. Khot
Principal

Agnel Charities' Fr. C. Rodrigues Institute of Technology
Vashi, Navi Mumbai 400 703, India

3rd Biennial International Conference on Nascent Technologies in Engineering (ICNTE2019)

Certificate

This is to certify that Dr/Mr/Ms. Chandresh Prasad

has presented paper titled. Intelligent News Aggregator and Validator

in the ICNTE2019 organized by Fr.C. Rodrigues Institute of Technology, Vashi in its premises in association with IEEE on January 4-5, 2019.

IEEE Advancing Technology for Humanity

ICNTE Connecting ideas.

Dr. Nitesh P. Yelve
Conference Chair

Dr. S.M. Khot
Principal

Agnel Charities' Fr. C. Rodrigues Institute of Technology
Vashi, Navi Mumbai 400 703, India

3rd Biennial International Conference on Nascent Technologies in Engineering (ICNTE2019)

Certificate

This is to certify that Dr/Mr/Ms. AMAANULLA KHATRI

has presented paper titled. INTELLIGENT NEWS AGGREGATOR AND VALIDATOR.

in the ICNTE2019 organized by Fr.C. Rodrigues Institute of Technology, Vashi in its premises in association with IEEE on January 4-5, 2019.

IEEE Advancing Technology for Humanity

ICNTE Connecting ideas.

Dr. Nitesh P. Yelve
Conference Chair

Dr. S.M. Khot
Principal

Agnel Charities' Fr. C. Rodrigues Institute of Technology
Vashi, Navi Mumbai 400 703, India

3rd Biennial International Conference on Nascent Technologies in Engineering (ICNTE2019)

Certificate

This is to certify that Dr/Mr/Ms. Irfan Khan

has presented paper titled... Intelligent News Aggregator and Validator

in the ICNTE2019 organized by Fr.C. Rodrigues Institute of Technology, Vashi in its premises in association with IEEE on January 4-5, 2019.

IEEE Advancing Technology for Humanity

ICNTE Connecting ideas.

Dr. Nitesh P. Yelve
Conference Chair

Dr. S.M. Khot
Principal