International Journal on Applications in Information and Communication Engineering
Volume 9 : Issue 1 : Feb 2023, pp 1 – 11 www.aetsjournal.com
ISSN (Online) : 2394 - 6237

A SOCIAL MEDIA FAKE NEWS DETECTION ALGORITHM USING MACHINE LEARNING

[1]D.KALEESWARAN, [2]SANTHOSH KUMAR S P, [3]HARISH R, 4SAMSUNDAR, 5MOHAMMED SULTAN
[1,2]Assistant Professor, [3,4,5]UG Scholar
Department of Information Technology,
Rathinam Technical Campus, Eachanari, Coimbatore, India

Abstract: - Using social media for news consumption has pros and downsides. Social media is used to find and consume news since it is free, easy to use, and delivers information quickly. On the other hand, it makes it easier for low-quality information that is deliberately incorrect to proliferate widely. The detection of bogus news on social media has thus become a major topic recently. It is vital to classify news as fake or not because false information spreads more quickly than true information. So, it is crucial to look into news authentication techniques. Word embedding (WE) over linguistic data is used to detect bogus news using machine learning classification. In the first stage, linguistic features are used to preprocess the data set and evaluate the accuracy of the news reports. At the second stage, several classification algorithms are used to combine the linguistic feature sets utilising Word Embedding approaches. To further support its methodology, which combines many data sets to produce an objective categorization result, this study painstakingly creates a brand-new Indian News data collection of over 55,000 items. With a 96.03% accuracy rate, the Indian News model classifies news as real or fake, outperforming Linear SVM TF-IDF representations and Random Forest models by 1.31% and 4.25%, respectively.

Key words: Fake News Detection, Social Media, Machine Learning, SVM, Naïve bayes.

I. INTRODUCTION

Fake news are low-quality information with purposefully false data, propagated by individuals or bots that deliberately manipulate message for tattle or political plans. It’s a false or misleading information presented as news. It often has the aim of damaging the reputation of a person or influence people’s views or making money through advertising revenue.

Most of the smart phone users prefer to read the news via social media over internet. The news websites are publishing the news and provide the source of authentication. The question is how to authenticate the news and articles which are circulated among social media like WhatsApp groups, Facebook Pages, Twitter and other micro blogs & social networking sites. It is harmful for the society to believe on the rumours and pretend to be a news. The need of an hour is to stop the rumours especially in the developing countries like India, and focus on the correct, authenticated news articles.

Fake news detection on social media presents unique characteristics and challenges that make existing detection algorithms from traditional news media ineffective or not applicable.

First, fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content; therefore, we need to include auxiliary information, such as user social engagements on social media, to help make a determination.

Second, exploiting this auxiliary information is challenging in and of itself as users’ social engagements with fake news produce data that is big, incomplete, unstructured, and noisy.

Fake news detection is made to stop the rumors that are being spread through the various platforms whether it be social media or messaging platforms, this is done to stop spreading fake news which leads to activities like mob lynching, this has been a great reason motivating us to work on this project. We have been continuously seeing various news of mob lynching that leads to the murder of an individual; fake news detection works on the objective of detecting this fake news and stopping activities like this thereby protecting the society from these unwanted acts of violence.

1.1 RELEVANCE OF THE PROJECT

A fake news classification system using different feature extraction methods and different classification algorithms like Support Vector Machine, Logistic Regression, Gradient Boosting, XG-BOOST, Decision Tree, Random Forest and the best algorithm we are going to use it in predicting the news as fake or real. In order to create a real time application, the algorithm should be fed with the most recent data.

Data is of different sizes so that should be properly cleaned to get better results. So we are using different algorithms and feature extraction methods like Bag of words model and Word embedding model to get the best result.

1.2 PROJECT OBJECTIVE

The main objective is to detect the fake news, which is a classic text classification problem with a straight forward proposition. It is needed to build a model that can differentiate between "Real" news and "Fake" news.

To achieve our goal of developing machine learning model to classify news as fake or real, we need perform following tasks in the same order as stated.

-Data Collection and Analysis

-Preprocessing the data

-Text feature extraction

-Using different classification algorithms

-Taking the best classification algorithm and feature extraction method

-Classifying the news as fake or real.

-Deploying the model.

1.4 SCOPE OF THE PROJECT

In future works, we intend to use highly sophisticated classifying approach, like deep learning with sentiment analysis also and consider many text features like publisher, urls etc., which may increase the accuracy of the classification of news as fake or real. Automatic fake news detection may be done using the latest news and training the model regularly to get the best results. So this can be used as a filter to upload the news.

2 SYSTEM ANALYSIS

2.1 Existing system:

This paper "WELFake: Word Embedding Over Linguistic Features for Fake News Detection" demonstrates a model and the methodology for fake news detection. With the help of Machine learning and natural language processing, author tried to aggregate the news and later determine whether the news is real or fake using Support Vector Machine. The results of the proposed model is compared with existing models. The proposed model is working well and defining the correctness of results upto 93.6% of accuracy**.**

Experimental results show that the WELFake model produces a high 91.73% accuracy on the WELFake data set. To further analyze its advantage we compared it with two state-of-the art works and found out that it improves the overall accuracy by 1.31% compared to BERT and 4.25% compared to CNN models. The proposed WELFake model also improved the accuracy by up to 10% on the McIntire and BuzzFeed data sets.

2.1.1 DISADVANTAGES OF EXISTING SYSTEM:

1. Now a days it takes much time to analyze trending in social media.
2. There lacks a performance evaluation of existing machine learning-based streaming fake news detection methods.
3. High level false positive rate.
4. Time complexity can be reduced.

2.2 PROPOSED SYSTEM:

The proposed system aims at utilizing the data collected from Kaggle website where real news and fake news were combined into one dataset. With different features and classification algorithms we are going to classify the news as fake or real and the algorithm with the feature which gives us the best result with that feature extraction method and that algorithm we are going to predict the news as fake or real. The system aims to investigate the utility of linguistic features for detecting fake news. We take supervised approaches like different classification algorithms such as Logistic Regression, SVM, Decision Tree, Random Forest etc., to the problem statement and ignoring attributes like the sources of the news, whether it was reported online or in print, etc. and instead focus only the content matter being reported. We aim to use different machine learning algorithms and determine the best way to classify news.

2.2.1 ADVANTAGES OF PROPOSED SYSTEM:

1. In this proposed system, the usage of multiple classification algorithms with different feature extraction methods, We could easily compare the accuracy by considering time and space complexity.

2. This resolves the issue of users being forged of fake news and could get and authenticated form of news. This would be ensured by storing the classified user information in the database.
3. Time complexity can be reduced.
4. Reduce the false positive rate and improve the accuracy.

3 System study

3.3.1 Feasibility Study:

During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the user. For feasibility analysis, some understanding of the major requirements for the system is essential. Two key considerations involved in the feasibility analysis are

- Economical feasibility
- Technical feasibility
- Operational feasibility

I. 3.3.1.1 Economical Feasibility:

Cost:

It is one of the economic analysis method used for evaluating the effectiveness of the system, commonly known as cost benefit analysis. The proposed project's cost and benefits are evaluated. Tangible costs include fixed and variable costs, while tangible benefits include cost savings, increased revenue, and increased profit.

II. Time:

The time required for generating an overall process comparatively very less than in the existing system. Here, all the works done through the software, it saves the time. When the user registering for the service provided it does not take much time and the common user can just login into the system using the email id and password.

III. 3.3.1.3 Operational feasibility:

The ability, desire, and willingness of the stakeholders to use, support, and operate the proposed computer information system. The stakeholders include management, employees, customers, and suppliers. The stakeholders are interested in systems that are easy to operate, make few, if any, errors, produce the desired information, and fall within the objectives of the organization.

2 SOFTWARE DESCRIPTION

2.1 Front end:

The part of a website that the user interacts directly is termed as front end. It is also referred to as the ‘client side’ of the application.

HTML: HTML is used to create and save web document. E.g. No

CSS : (Cascading Style Sheets) Create attractive Layout

JavaScript: it is a programming language, commonly use with web browsers.

2.1.1 HTML

HTML stands for Hyper Text Mark up Language. It is used to design the front end portion of web pages using mark up language. It acts as a skeleton for a website since it is used to make the structure of a website.

1. Publish online documents with headings, text, tables, lists, photos, etc.
2. Retrieve online information via hypertext links, at the click of a button.
3. Design forms for conducting transactions with remote services, for use in searching for information, making reservations, ordering products, etc.

Include spread-sheets, video clips, sound clips, and other applications directly in their documents. With HTML, authors describe the structure of pages using markup. The elements of the language label pieces of content such as “paragraph,” “list,” “table,” and so on.

2.2.2 CSS

Cascading Style Sheets fondly referred to as CSS is a simply designed language intended to simplify the process of making web pages presentable. It is used to style our website.

It allows one to adapt the presentation to different types of devices, such as large screens, small screens, or printers. CSS is independent of HTML and can be used with any XML-based markup language.

The separation of HTML from CSS makes it easier to maintain sites, share style sheets across pages, and tailor pages to different environments. This is referred to as the separation of structure.

2.2.3 JAVASCRIPT

JavaScript is a scripting language used to provide a dynamic behavior to our website.

JavaScript is the Programming language of HTML and web. JavaScript is one of the 3 languages all web developer must learn:

- HTML to define the content of web pages.
- CSS to specify the layout of web pages.
- JavaScript to program the behaviour of web pages

  JavaScript and Java are completely different languages, both in concept and design.

2.2.4 BOOTSTRAP

Bootstrap is a free and open-source tool collection for creating responsive websites and web applications. It is the most popular CSS framework for developing responsive, mobile-first websites. Nowadays, the websites are perfect for all the browsers (IE, Firefox, and Chrome) and for all sizes of screens (Desktop, Tablets, Phablets, and Phones).

2.3 BACK END:

Backend is the server side of a website. It is the part of the website that users cannot see and interact. It is the portion of software that does not come in direct contact with the users. It is used to store and arrange data.

Python: Python is a programming language that lets you work quickly and integrate systems more efficiently.

MYSQL: MYSQL is a database, widely used for accessing querying, updating, and managing data in databases.

2.3.1 PYTHON

Python is a general-purpose programming language used in web development to create dynamic websites using frameworks like Flask, Django, and Pyramid. For the most part, Python runs on Google's Apps Engine.

It is used in backend development while its frameworks are used in frontend development. Python is therefore an ideal backend language due to its simplicity and consistency, where the developers are able to write reliable systems with a vast set of libraries belonging to Machine Learning, Keras, TensorFlow and Scikit-learn. Python's extensive set of libraries and frameworks can be extremely useful and time-saving, which results in quicker turnover times and more productivity.

2.3.2 MYSQL

To work with data in a database, you must use a set of commands and statements (language) defined by the DBMS software. There are several different languages that can be used with relational databases; the most common is SQL. Both the American National Standards Institute (ANSI) and the International Standards Organization (ISO) have defined standards for SQL. Most modern DBMS products support the Entry Level of SQL-92, the latest SQL standard (published in 1992).

Microsoft SQL Server supports a set of features that result in the following benefits. SQL Server includes a set of administrative and development tools that improve your ability to install, deploy, manage, and use SQL Server across several sites. SQL Server integrates with e-mail, the Internet, and Windows.

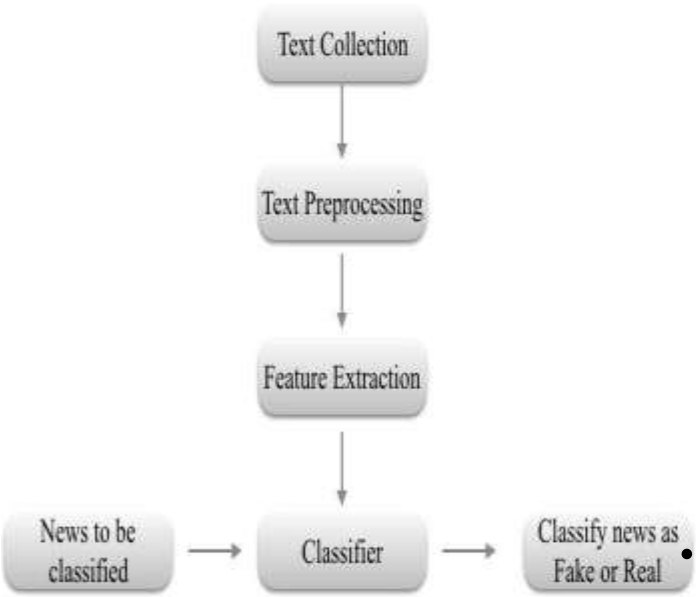
A database in Microsoft SQL Server consists of a collection of tables that contain data, and other objects, such as views, indexes, stored procedures, and triggers, defined to support activities performed with the data.

2 SYSTEM DESIGN

2.1 System Architecture

A system architecture diagram would be used to show the relationship between different components. Usually they are created for systems which include hardware and software and these are represented in the diagram to show the interaction between them. However, it can also be created for web application. For a web application the system architecture design

would include components such as, database, application server, web server, internet, browser etc. □ Converting all letters to lower case Browser can be connected into server contains the Pages, CSS, images.



5.1.1. Text Collection

The dataset was taken from Kaggle the content and metadata has been extracted from 244 web sites that have been considered to be associated with fake news by the BS Detector Chrome Extension by Daniel Sieradski. It consists of almost 13000 posts over a period of 30 days. Research on this dataset using language processing tools has already been carried out by Kaggle users. The dataset was generated by Andrew Thompson to create document term matrices using the articles and analyse connections between articles using common political affiliations, medium or subject matter. It contains articles from top 15 American publications and the articles were mostly published between the years of 2016 and 2017. It consists of around 150000 articles that were collected by scraping news website homepages and RSS feeds. However, we will randomly select only 13000 articles from this dataset and merge it with the fake news dataset for more accurate predictions and for avoiding a skewed dataset.

$$idf_j = log\left[\frac{n}{df_j}\right]$$

IV. 5.1.2. Text Preprocessing

After a text is obtained, we start with text pre processing. Text pre processing includes:

- Removing punctuations, accent marks
- Removing white spaces
- Removing stop words

Feature Extraction

Text needs to be converted into numbers before it is used with a machine learning algorithm. For classification of documents, documents are taken as input and a class label is generated as output by the predictive algorithm. The documents need to be converted into fixed-length vectors of numbers for the algorithm to take them as input.The input for the machine learning algorithm are the words encoded as integers or floating point values.

V. Bag of Words (BOW)

We make the list of unique words in the text corpus called vocabulary. Then we can represent each sentence or document as a vector with each word represented as 1 for present and 0 for absent from the vocabulary.

- Count Vectorizer

Count Vectorizer generates an encoded vector that contains the length of the }entire vocabulary coupled with the frequency of each word by which it appears in the document.

- Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency (TF) = (Number of times term t appears in a document)/(Number of terms in the document)

$$TF_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{i,j}}$$

Inverse Document Frequency (IDF) = log(N/n), where, N is the number of documents and n is the number of documents a term t has appeared in. The IDF of a rare word is high, whereas the IDF of a frequent word is likely to be low. Thus having the effect of highlighting words that are distinct.

We calculate TF-IDF value of a term as = TF * IDF

- Word Embedding

It is a representation of text where words that have the same meaning have a similar representation. In other words it represents words in a coordinate system where related words, based on a corpus of relationships, are placed closer together. Word embeddings are in fact a class of techniques where individual words are represented as real-valued vectors in a

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

predefined vector space. Each word is mapped to one vector.Each word is represented by a real-valued vector.

5.1.3 CLASSIFIER

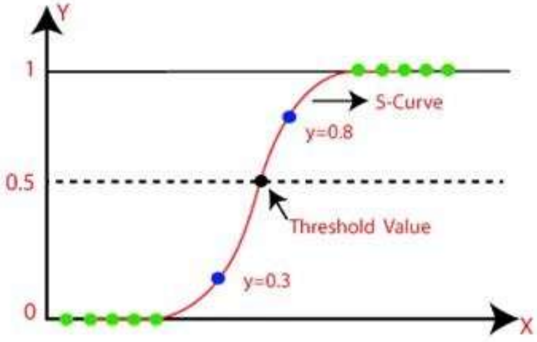
The feature vectors are send to the classifier to classify the news as fake or not.

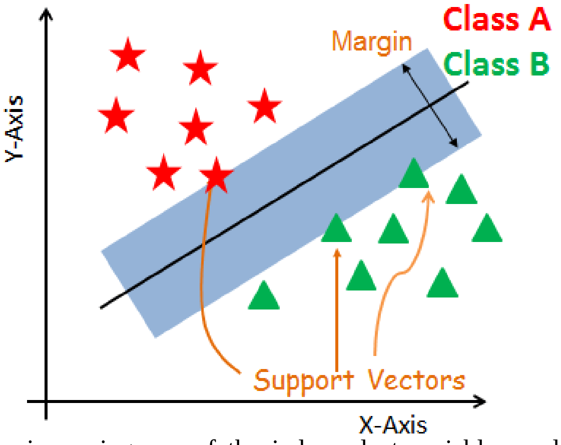
5.1.3.1 SUPPORT VECTOR MACHINE

A support vector machine is in a way a binary classifier since the model works by generating a hyperplane that is used to separate the training data as far as possible. The support vector machine performs well since there is an extremely high number of features in a text classification problem but generally requires a lot of tuning and is memory intensive. Once we have labelled training data (supervised learning) the algorithm generates the best possible hyperplane which categorizes new data automatically. In a two dimensional space this hyperplane would be a line dividing a plane in two parts with each class lying on either side.

5.1.3.2 Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.





The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.

VI. FIGURE 5.3 LOGISTIC REGRESSION

5.1.3.3 Decision Tree

Decision Tree is one of Supervised Machine Learning (output label is given) technique where the data is split consecutively through a definite parameter.

In a decision tree every node speaks to a feature(attribute), each link(branch) speaks to a decision(rule) and each leaf speaks to an outcome(categorical or continuous value).The entire algorithm is to make a tree for the entire information and process a solitary result at each leaf(or limit error in each leaf). The process of text classification using decision tree marks internal node as terms and the branches withdrawing from them as derived weight, and relative class labels are represented by leaf node. Decision tree utilizes query structure throughout the path of the tree classifying the document from root until it reaches a definite leaf node. In memory decision tree construction, majority of training data will not fit and results inefficient due to swapping the training tuples. This is dealt in as FDT to deal with the multiclass record which lessens the induction cost. A symbolic rule induction system based on decision tree is presented to improve text classification to implement multiclass classification.

5.1.3.4 RANDOM FOREST

As the name suggests, Random Forest algorithm generates the forest with a number of decision trees. So it is the collection of decision trees. Decision trees are attractive classifiers among others because of their high execution speed. Based on random samples from the database a random forest classifier averages multiple decision trees. Generally, the more trees in the forest is the sign that forest is robust. Similarly in the random forest classifier, high accuracy is obtained by higher the number of trees in the forest. While concurrently creating a tree with decision nodes, a decision tree breaks the dataset down into smaller subsets. The decision root node is selected through highest information gain and leaf nodes based on a pure subset for each iteration simultaneously. Calculation of Information Gain (IG) requires impurity measure (Entropy) of that node. There are various indices to measure the degree of impurity. A leaf node represents a category or pure subset. The trees in a random forest are created under random data so there might be chances to be lack meaning and noisy. In order to make a model with low variance random forest averages these trees. The irrelevant trees drop each other out and the staying meaningful trees yield the final result.

5.1.3.5 GRADIENT BOOSTING

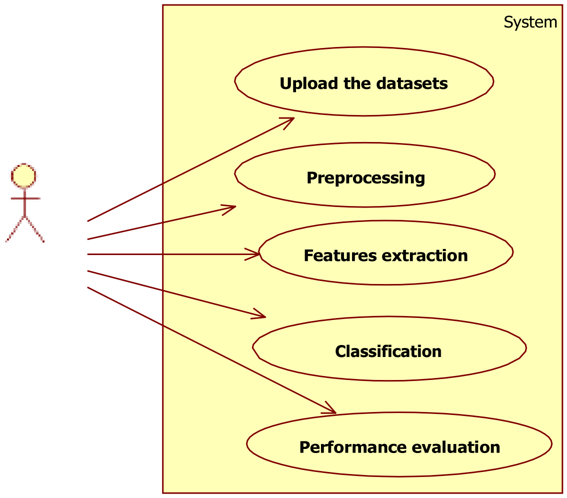
Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

5.1.3.6 XG-BOOST

This is an ensemble method that seeks to create a strong classifier (model) based on “weak” classifiers. In this context, weak and strong refer to a measure of how correlated are the learners to the actual target variable. By adding models on top of each other iteratively, the errors of the previous model are corrected by the next predictor, until the training data is accurately predicted or reproduced by the model.

Among these classifiers and feature extraction method the best classification algorithm and feature extraction method is used to classify the news as fake or real.

Use Case Diagram:
A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well.

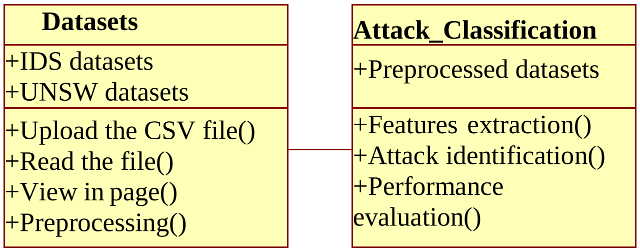


I. FIGURE: 5.4 USECASE DIAGRAM

5.3 Class Diagram

The class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing and documenting different aspects of a system but also for construct in executable code of the software application.

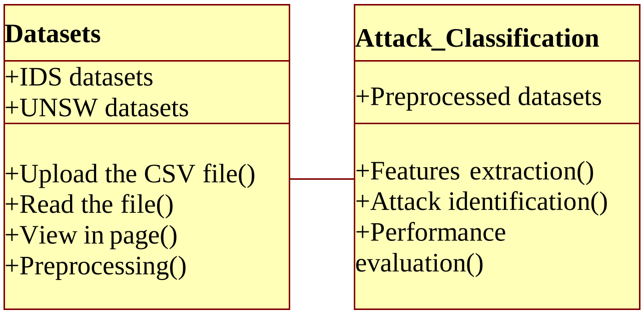
The class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modelling of object oriented systems because they are the only UML diagrams which can be mapped directly with object oriented languages. The class diagram shows a collection of classes, interfaces, associations, collaborations and constraints. It is also known as a structural diagram. The purpose of the class diagram is to model the static view of an application.



I. FIGURE 5.5 CLASS DIAGRAM

5.4 ACTIVITY DIAGRAM

Activity diagram is another important diagram in UML to describe dynamic aspects of the system. Activity diagram is basically a flow chart to represent the flow form one activity to another activity. The activity can be described as an operation of the system. So the control flow is drawn from one operation to another. This flow can be sequential, branched or concurrent. Activity diagrams deals with all type of flow control by using different elements like fork, join etc. The basic purposes of activity diagrams are similar to other four diagrams. It captures the dynamic behaviour of the



I. FIGURE 5.5 CLASS DIAGRAM

5.4 ACTIVITY DIAGRAM

Activity diagram is another important diagram in UML to describe dynamic aspects of the system. Activity diagram is basically a flow chart to represent the flow form one activity to another activity. The activity can be described as an operation of the system. So the control flow is drawn from one operation to another. This flow can be sequential, branched or concurrent. Activity diagrams deals with all type of flow control by using different elements like fork, join etc. The basic purposes of activity diagrams are similar to other four diagrams. It captures the dynamic behaviour of the

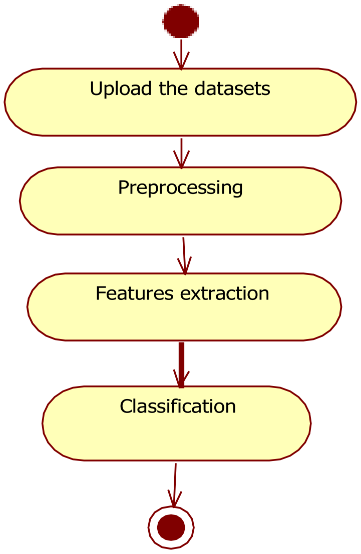
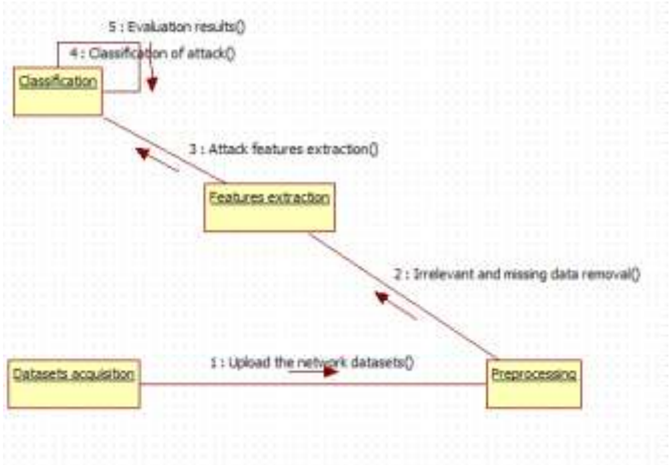


Figure 5.6 Activity diagram

5.6 SEQUENCE DIAGRAM

A sequence diagram in a Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of


5 : Evaluation results()
4 : Classification of attack()
Classification
3 : Attack features extraction()
Features extraction
2 : Irrelevant and missing data removal()
1 : Upload the network datasets()
Datasets acquisition
Preprocessing

messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams typically are associated with use case realizations in the Logical View of the system under development. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing

Diagrams.

2 PROJECT DESCRIPTION

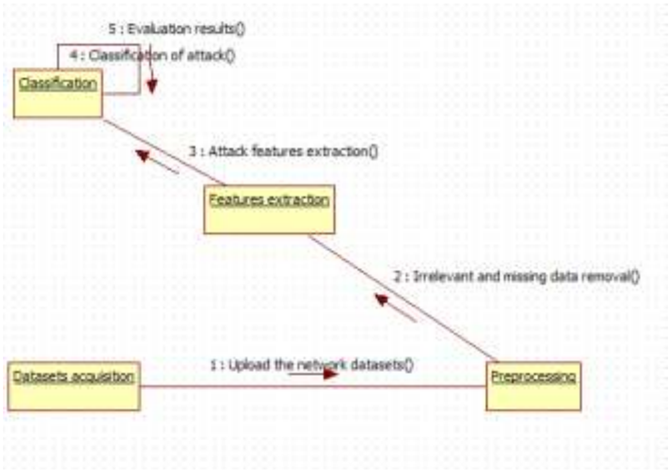
Figure 5.8 Collaboration diagram

2.1 Module Design

This project consists of modules. Each and every module performs a particular work. The modules are

5.8 COLLABORATION DIAGRAM

A collaboration diagram, also called a communication diagram or interaction diagram, is an illustration of the relationships and interactions among software objects in the Unified Modeling Language (UML). The concept is more than a decade old although it has been refined as modeling paradigms have evolved. A collaboration diagram resembles a flow chart that portrays the roles, functionality and behavior of individual objects as well as the overall operation of the system in real time. Objects are shown as rectangles with naming labels inside. These labels are preceded by colons and may be underlined. The relationships between the objects are shown aslines connecting the rectangles


5 : Evaluation results()
4 : Classification of attack()
Classification
3 : Attack features extraction()
Features extraction
2 : Irrelevant and missing data removal()
Datasets acquisition
1 : Upload the network datasets()
Preprocessing

2.4 Module description

6.1.1 Datasets Acquisition

In this module, we can upload the news datasets in the form of CSV file from the website of Kaggle.

6.1.2 PREPROCESSING

Dataset Acquisition Module

Data pre-processing is an important step in the [data mining] process. The phrase "garbag in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. In this module, eliminate the irrelevant and missing values in uploaded datasets.

6.1.3 FEATURES EXTRACTION

Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Many machine learning practitioners believe that properly optimized feature extraction is the key to effective model construction. Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. In this module, we can select the many attributes from pre-processed datasets

6.1.4 CLASSIFICATION

As the demand for data collecting and analytics grows, so does the cost of doing so. It's difficult to distinguish false news from news sourced from multiple social media platforms because news can take several forms, including texts, hyperlinks, photographs, and videos.

Various machine techniques are employed to detect yellow news or fake news in this module.

We will categorise the news as fake or real using several feature extraction methods and classification algorithms, and we will predict the news as fake or real using the algorithm with the feature that gives us the best result using that feature extraction method and algorithm.

6.1.5 PERFORMANCE EVALUATION

In this module, performance can be evaluated in terms of accuracy rate. Proposed work provide improved accuracy rate than the existing systems.

3 TESTING

This chapter gives an overview of the various types of testing incorporated during the entire duration of the project.

3.1 UNIT TESTING:

SVM Linear, Decision Tree, Logistic Regression Classifier,Random forest,XG-BOOST, Gradient Boosting algorithms were separately tested to see if they were able to give the accuracy better than the existing system. These were tested separately so that, we could compare between each other.

3.2 SYSTEM TESTING:

Under System Testing technique, the entire system is tested as per the requirements. It is a Black-box type testing that is based on overall requirement specifications and covers all the combined parts of a system. Here we tested if the end integrated code could run on any system, we saw that the integrated code can run on any system having python version 3.6 or more, and we never faced any error.

3.3 USABILITY TESTING:

This project could be easy for python and data science programmer, not meant for general purpose. Application is usable for data science engineers to pick the model for future research in the area of fake news classification.

4 CONCLUSION AND FUTURE SCOPE

4.1 Conclusion

In this project three different feature extraction methods like Count Vectorizer,TFIDF Vectorizer,Word Embedding has been used.And also different classification algorithms like Linear SVC,Logistic Regression Classifier,Decision Tree Classier,Random Forest Classifier,XG-BOOST Classifier,Gradient Boosting Classifier have been used to classify the news as fake or real. By using the classification algorithms we got highest accuracy with SVM Linear classification algorithm and with TF-IDF feature extraction with 0.94 accuracy.Even though we got the same accuracy with Neural Network with Count Vectorizer, Neural Networks and take more time to train and its complex so we used Linear SVC which is not so complex and takes less time to compute.

4.2 FUTURE SCOPE

In future we can also use deep learning methods and sentiment analysis to classify the news as fake or real which may get high accuracy and we can extract further useful text like publication of the news, url domain etc., We can use more data for training purposes - In machine learning problems usually availability of more data significantly improves the performance of a learning algorithm. Dataset with larger number of news articles from different sources would be of a great help for the learning process as news from different sources will involve larger vocabulary and greater content.

9. RESULTS

Classifier	Word embedding	Count Vectorizer	TF-IDF
SVM LINEAR	0.87	0.91	**0.94**
LOGISTIC REGRESSIO N	0.87	0.93	0.93
DECISION TREE	0.73	0.82	0.82
RANDOM FOREST	0.84	0.88	0.90
XG-BOOST	0.83	0.89	0.89
GRADIENT BOOSTING	0.83	0.89	0.90
NEURAL NETWORK	0.90	0.94	0.93

10. REFERENCES

[1] Ahmed .H, Traore. I, and Saad.S, "Detection of online fake news using N-gram analysis and machine learning techniques," in Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, vol. 10618. Cham, Switzerland: Springer, Oct. 2017, pp. 127–138.

[2] Buntain.C and Golbeck.J, "Automatically identifying fake news in popular Twitter threads," in Proc. IEEE Int. Conf. Smart Cloud (SmartCloud), New York, Nov. 2017, pp. 208–215.

[3] Burgoon J. K., Blair J, T. Qin, and J. Nunamaker, "Detecting deception through linguistic analysis," in Proc. 1st NSF/NIJ Conf. Intell. Secur. Inform., Berlin, Germany: Springer, May 2003, pp. 91–101

[4] Gravanis.G, Vakali.A, Diamantaras.K, and Karadais.P, "Behind the cues: A benchmarking study for fake news detection," Expert Syst. Appl., vol. 128, pp. 201–213, Aug. 2019.

[5] Jiawei Zhang; Bowen Dong; Philip S. Yu, FakeDetector: Effective Fake News Detection with Deep Diffusive Neural Network ,2020 IEEE 36th International Conference on Data Engineering (ICDE), Apr.2020

Pawan Kumar Verma, Member, IEEE, Prateek Agrawal, Ivone Amorim, and Radu Prodan, Member, IEEE "WELFake: Word Embedding Over Linguistic Features for Fake News Detection", IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, VOL. 8, NO. 4, AUGUST 2021

[6] Pérez-Rosas V., Kleinberg.B, Lefevre A., and Mihalcea R., "Automatic detection of fake news," in Proc. 27th Int. Conf. Comput. Linguistics, Santa Fe, NM, USA: Association for Computational Linguistics, Aug. 2018, pp. 3391–3401.

[7] Shu.K, Wang.S, and Liu.H, "Exploiting Tri-relationship for fake news detection," Dec. 2018, arXiv:1712.07709v1. [Online].

[8] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, Junzhou Huang "Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks".

[9] Uma Sharma, Sidarth Saran, Shankar M. Patil Department of Information Technology Bharati Vidyapeeth College of Engineering Navi Mumbai, "India Fake News Detection using Machine Learning Algorithms", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org NTASU - 2020 Conference Proceedings