IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Website Legitimacy and Fact Checking

Manish Chauhan[1] , Naresh Metre[2] , Abhishek Mhatre[3] ,
Prof. Shilpa M. Satre[4]

[1]*Manish Chauhan, Department of Information Technology, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India.*

[2]*Naresh Metre, Department of Information Technology, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India.*

[3]*Abhishek Mhatre, Department of Information Technology, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India.*

[4]*Prof. Shilpa M. Satre, Dept of Information Technology, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India.*

ABSTRACT: In today's world, technology plays an important role in individual's lives. It has evolved into a valuable and practical tool for facilitating public transactions such as e-banking and e-commerce. As a result, users have come to believe that it is safe to give their personal information to the Internet. As a result, the security thieves who have begun to target this data have turned into a major security issue. One of these issues is the use of phishing websites. A malicious URL is a common and real cyber security threat. Innocent users who visit such websites risk becoming victims of a variety of scams, including financial loss and the theft of personal information. It is critical to recognize and react to such threats as soon as possible. This project proposes a system which allows user to check the legitimacy of the website and to detect the URL is fake or real. It also helps in detecting the fake information.

KEYWORDS: Phishing, Legitimacy, Support Vector Machine, Fact Checking, Random Forest

I. INTRODUCTION

Phishing is an effective method of deceiving people, either by giving the impression that the site is legitimate or by displaying greedy tactics. Nowadays, it is affecting both financial and individual organizations a lot. Attackers use a variety of methods to steal information, including email, advertising, and fake websites. Phishing has been one of the most common cyber-attacks. Phishing attacks mainly utilize the social engineering and technology deception to obtain user privacy information. The most common way of phishing attacks is to send an illegal link to the user and induce the users to click. The users are then duped into providing personal information without their knowledge. And Phishing websites also helps in spreading rumors or fake news. The issue of detecting false news, on the other hand, has its own unique characteristics that greatly increase the challenge. When fake news reaches a large number of social media users, it is compounded by an "echo chamber effect," in which users who agree with the fake news are more likely to share it

Nowadays, due to rapid growth and advancement in the upcoming technologies internet has becoming the vital and useful in numerous fields like E-Commerce, Finance, Business, Social Networks, etc [12]. One effective way to mitigate this type of problem is warning or preventing the users from clicking the link of phishing web pages. However, identifying whether a webpage is legitimate or phishing is a very challenging problem because phishing attack has semantics- based structure and mainly exploits the computer user's vulnerabilities. Several anti-phishing approaches have been developed, that they use blacklists, visual and machine learning-based techniques. Unfortunately, these approaches cannot prevent all types of phishing attacks, especially zero-day ones. Blacklist-based approach [3] is a well-known method to detect phishing websites. A list of phishing URLs and domain names is included. If a user clicks on a website, immediately its URL would be checked against the blacklist database. Current practices do a decent job of detecting known phishing websites, but there is a time delay before those phishing websites are added to the blacklist.

In this project we are developing a system in which we check the authenticity of the website or URL based on various hyperlink features [5] using Random Forest algorithm. There are many websites which are fake and resembles similar to the original website, we determine such types of malicious websites. Secondly we are developing a system in which can check the authenticity of an image and detects whether image is fake or real. By using Tesseract tool we are extracting the text from the image and checking the fact present on image is true or not. We are also developing a fact checking tool in our system, which will verify that text content is fake or not.

A) PROBLEM STATEMENT

Phishing is a major concern as it leads to many problems. Attackers are using a social engineering trick, which can be described as fraudsters that try to manipulate the user into giving them their personal information based on exploiting human vulnerabilities rather than software vulnerabilities. Phishing on the internet is becoming more popular by the day. It is an unauthorised attempt by attackers to obtain personal information such as bank account numbers, login ids, and passwords. Authentic links, such as to the real privacy policy and terms of service pages for the site they're imitating, are also included in spoof pages. To make the spoof site look more believable, these genuine links are mixed in with links to a fake phishing website. These types of websites are very difficult to identify, as it resembles exactly similar to the original one. The unauthorized websites leads to unauthorized or illegal fake content, which provides wrong information to the user. These websites contains various media files which may provide improper information or false news, which may get viral on social networking sites. Without understanding the source of the news, it is difficult to distinguish between real news and hoaxes at the current rate of news generation on social media. The widespread dissemination of false news has the potential to have highly damaging consequences for both individuals and society. As a result, detecting false news on social media has recently become an emerging research topic that is gaining a lot of attention.

B) OBJECTIVES

- The project aims to detect legitimacy of website.
- User will learn about factors that determines legitimacy of a website.
- Our aim is to make users avoid being attacked by phishing websites.
- Our aim is to detect fake information about covid-19, including which are spreading through information (text) on images.
- User will try to avoid misleading or harmful information about covid-19.

II. LITERATURE REVIEW

Several researchers suggested various methods for detecting phishing URLs. Some of them also held a list of previously identified phishing websites' domain names or IP addresses. Following are some references which highlight the literature review:

According to Mohammad Mehdi Yadollahi, Farzaneh Shoeleh [1], in this paper, a real- time anti-phishing system is proposed. Different types of discriminative features are used in this paper, including URL-based features, HTML-based features, statistical-based features, and NLP-based features.

The phishing attack is identified in [2] by examining the hyperlinks contained in the website's HTML source code. To detect phishing attacks, the proposed solution integrates a number of new outstanding hyperlink unique features. The proposed method categorises the hyperlink-specific features into 12 categories and uses these categories to train machine learning algorithms.

A system named Phishnet is proposed [3] where a blacklist of phishing URL was maintained. It will determine whether the IP address, hostname, or URL itself is on the blacklist.

According to K. Arvind, S. Govarthan, S. Kishore Kumar [4], the collected knowledge is classified into two groups using the Bayesian classification algorithm: fake and authentic. Reverse Dissemination process is done to back track the rumour to its source.

In [5] Abhijit Sharma proposed a system in which dataset of phishing and legitimate URLs are collected. Lexical features of these URLs are extracted. Feature selection method is used to find the important features only. This approach assigns a rank to each feature based on how well it contributes to the detection of phishing and non-phishing threats.

According to Zaitul Iradah Mahid, Selvakumar Manickam, Shankar Karuppayah [6], discuss various deception detection methods, which are divided into three categories: content-based, social context-based, and hybrid-based methods. These detection systems, which are only used in a few domains, predict news content manipulation with a high degree of accuracy.

In [7], an intelligent detection system based on Ensemble Voting Classifier is proposed to deal with both authentic and false news classification tasks. For detection, Naive Bayes, K-NN, SVM, Random Forest, Artificial Neural Network, Logistic Regression, Gradient Boosting, Ada Boosting, and other well-known machine-learning algorithms are used. In Ensemble Voting Classifier, the best three machine learning algorithms were used after cross-validation.

In [8] a system is proposed to detect the authenticity of image using metadata technique and error level analysis. It is basically used to detect the image that is Photoshopped or a gimped image. The system fails on images shared through Whatsapp, Google+ etc.

According to Aliya Begum, Srinivasu Badugu in [9], Studied different techniques for detecting malicious URL and discussed each and every technique with their merits and demerits. Deep learning approaches are a promising direction in detection of phishing URL.

According to Kuai Xu, Feng Wang, Haiyan Wang and Bo Yang [10], Here two approaches are used for the detection of fake news content i.e. Domain reputation and content understanding. Here system analyses both domain of the website and the fake content.

III. PROPOSED SYSTEM

In this project we are developing a system in which we check the authenticity of the website or URL. There are many websites which are fake and resembles similar to the original website, we determine such types of malicious websites. Secondly we are developing a system which can check the authenticity of an image and detects whether image is fake or real. We are also developing a fake information detection tool in our system, which will verify that text content about COVID-19 is fake or real.

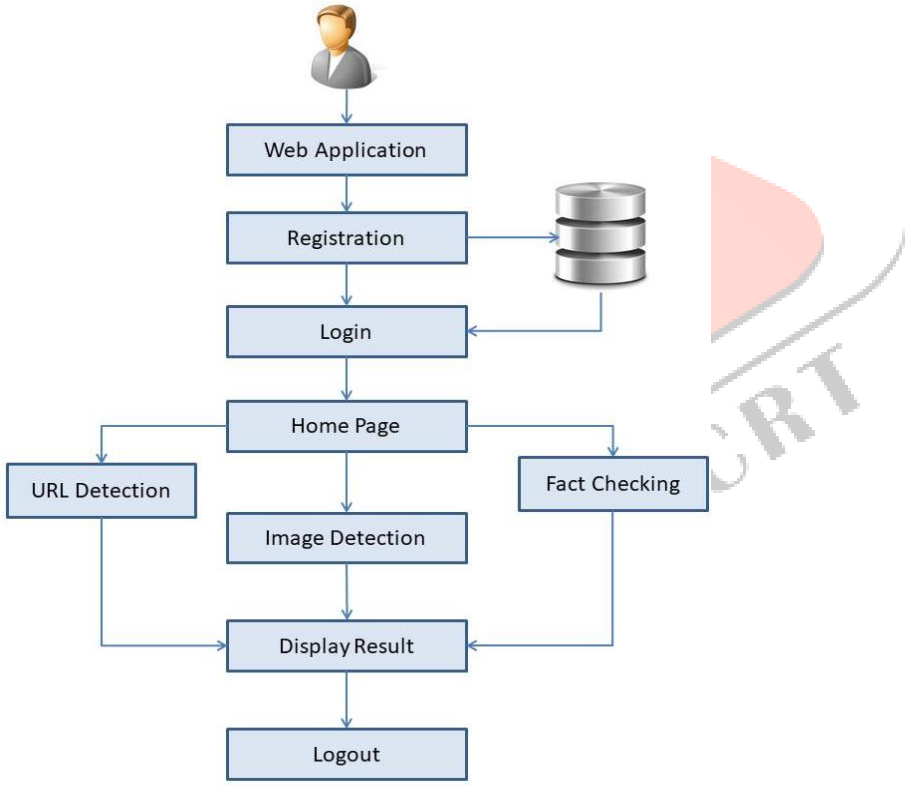

Fig 1: System Architecture

A) SYSTEM METHODOLOGY

Firstly, the user needs to register them in the system. If the user is already a registered user, then the user needs to log in to the system with the valid username and password to use the system. The authentic user will give input to the system, if the input is website or URL, and then the system will follow following steps to check the legitimacy of website:

- The System will extract various features, which will be responsible for finding the proper feature set.
- As the features are based on hyperlinks of the webpage, the website will be transformed into DOM tree, which will be used to extract hyperlink features.
- Then the system will select proper feature set.
- The system will have a constructed training, test dataset and model for training data by extraction of features from phishing or legitimate website and its hyperlinks.
- In the training phase, the Random Forest algorithm model applies selected features of the dataset to the input website or URL and its hyperlinks.
- In the testing phase, the Random Forest algorithm model determines whether the input URL is legitimate website or not, based on learning from the dataset.

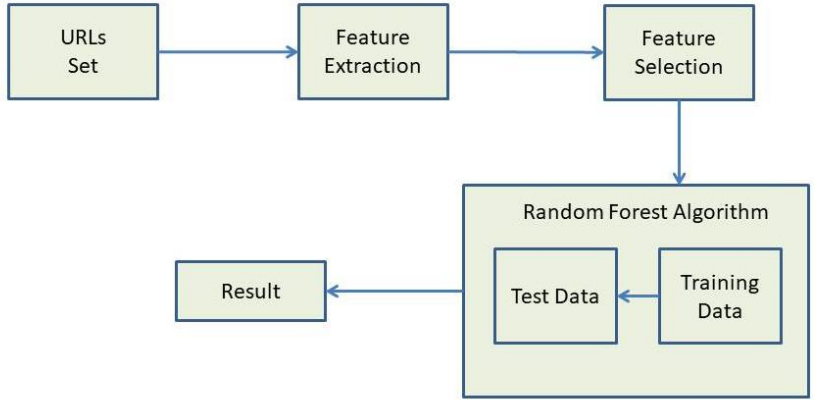

Fig 2: Block diagram of website authenticity process

If the input is image, then the system will follow following steps:

- First the system will extract text from image using Tesseract tool.
- The text extracted from image will be analysed, the result will be displayed as fake or real content.

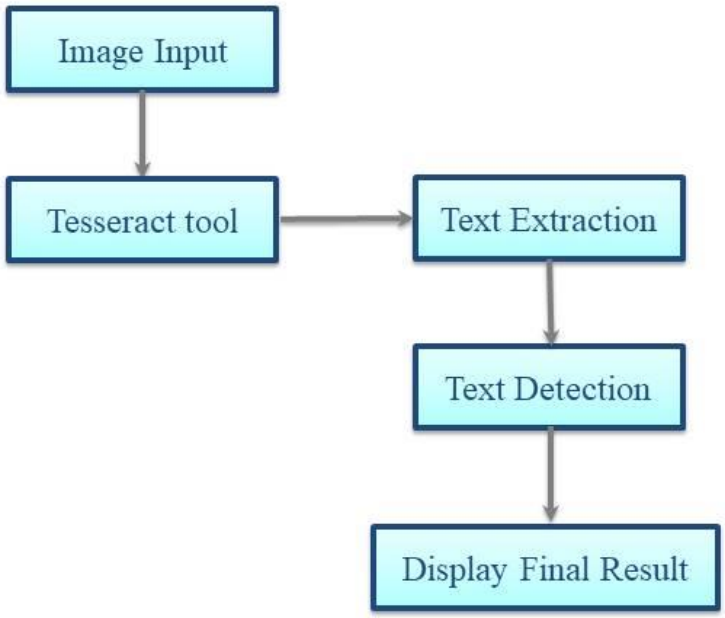

Fig 3: Block diagram of Image detection process

If system finds input as Text or information content, then the system will follow following steps:

- The extracted text will be trained in SVM algorithm model.
- The SVM algorithm model will be provided a related content queried from dataset.
- The extracted text will be tested in dataset.
- Then the output will show the results (classified as Fake and Real)

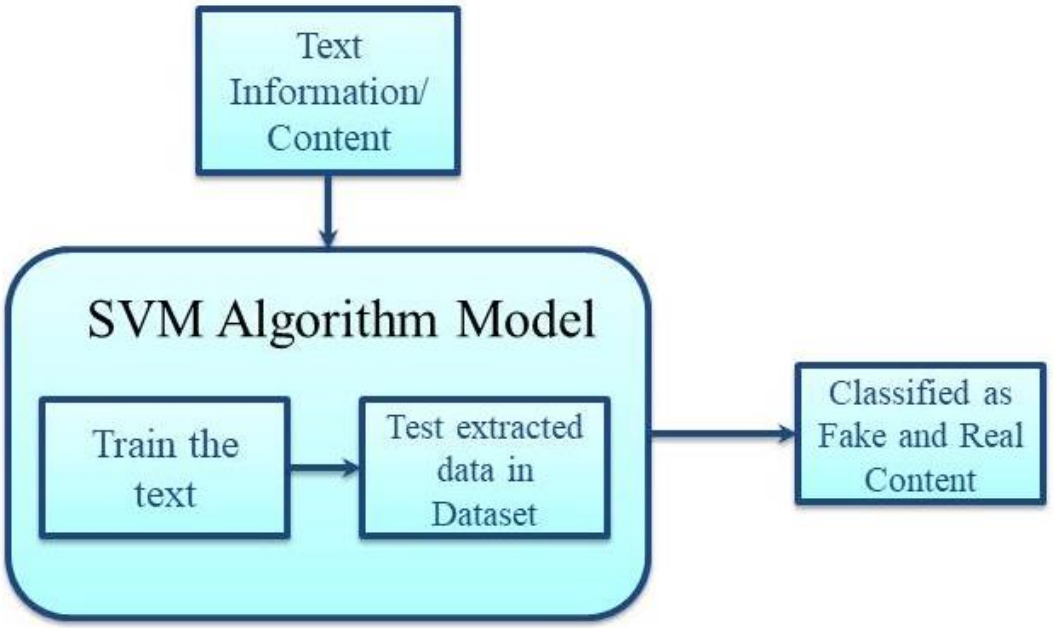

Fig 4: Block diagram of Text detection process

IV. RESULTS AND DISCUSSION

This project "Website Legitimacy and Fact Checking" proves to be an effective and efficient system which can be easily accessed by the user. It is a web based project, which is implemented in Django framework. The user can access the portal through registration process and can use the features provided by the system. User can check the authenticity of any URL in the system. System also offers to detect the authenticity of images and can check the information of COVID-19 which is circulated on social media.

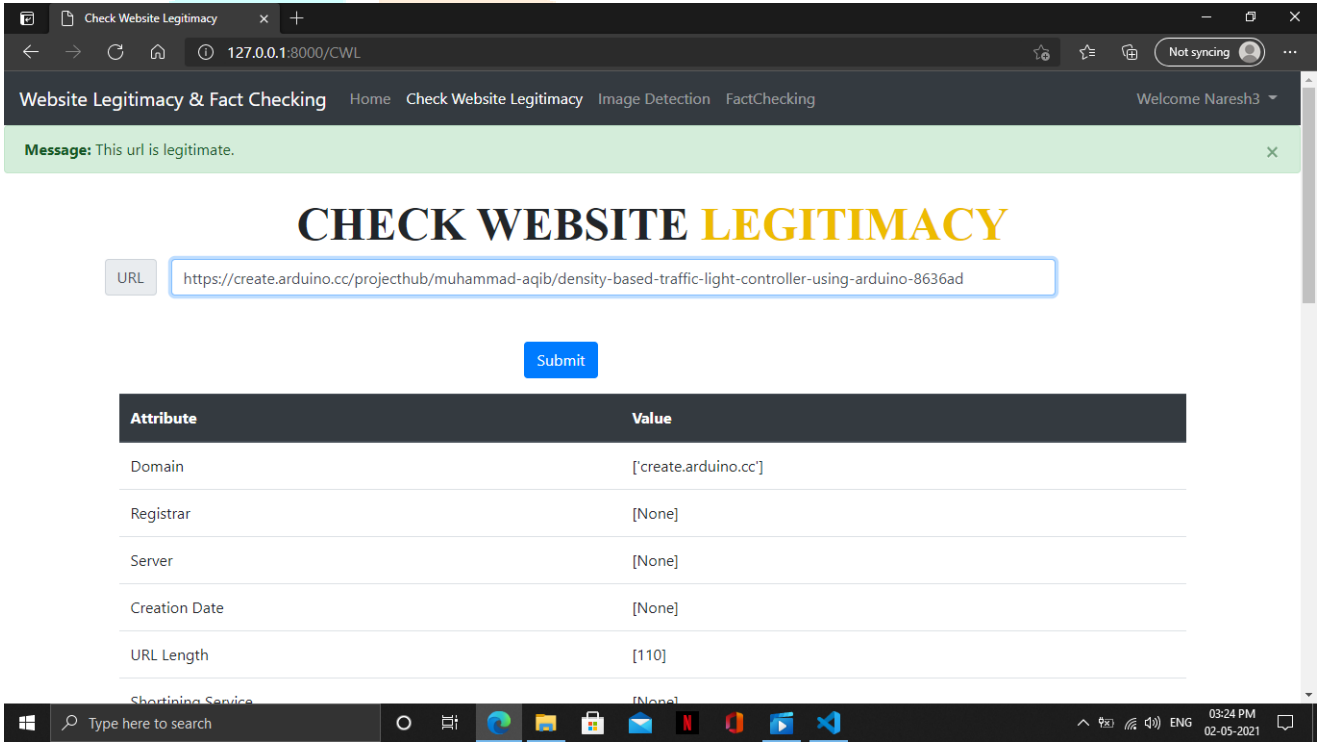

Fig 5: Website Detection process through website features

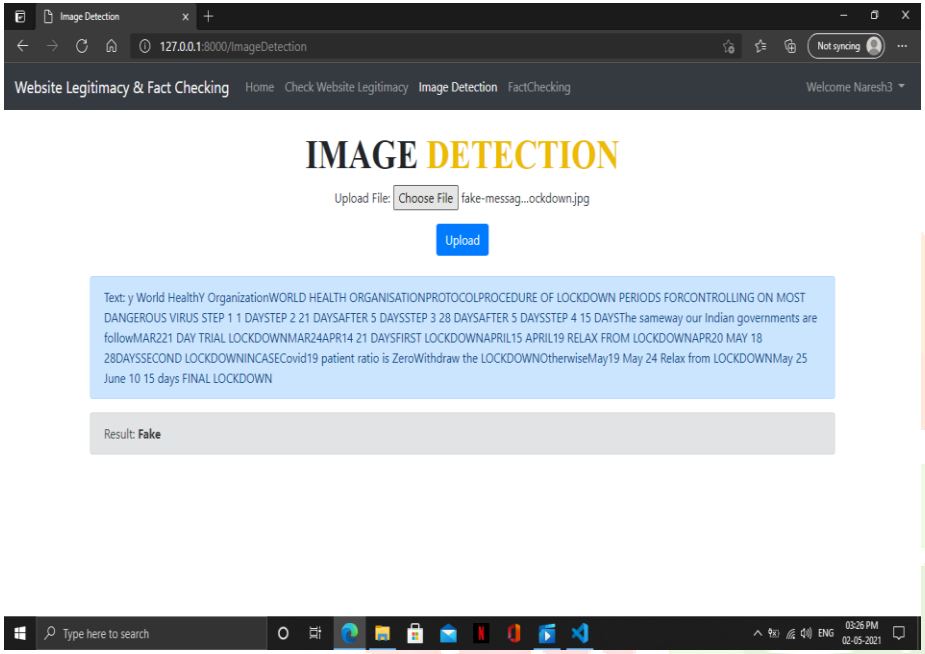

Fig 6: Text Extraction from Image and Image Detection process

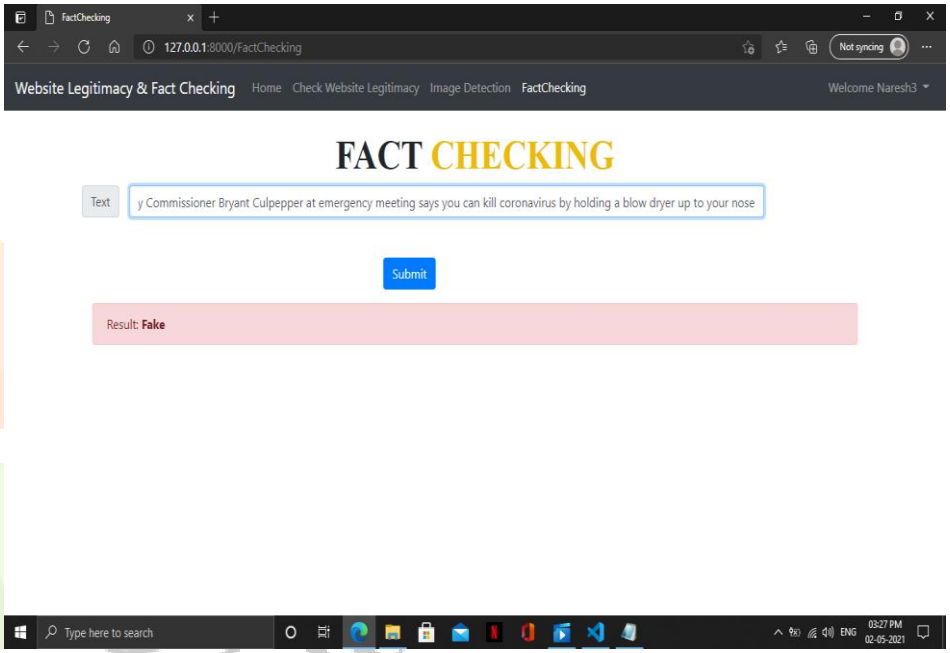

Fig 7: Fake news detection (COVID-19 News) system

A) ADVANTAGES

- System is able to detect misinformation or fake content related to COVID-19.
- It is able to detect fake URLs or malicious website.
- Avoid people to use fake duplicate websites.
- System is user friendly and easily operable.
- Results are reliable and accuracy level is efficient.

B) APPLICATIONS

- Anti-phishing browser can be developed using same technique.
- Can be used in social media platforms to avoid phishing attack.
- Social networking sites can use this technique to verify the information that is circulated on it.
- Image detection function can be used in many Image related websites like Shutterstocks, Pexels etc.

V. CONCLUSION

In this project, we have implemented a system that checks legitimacy of a website. It will analyse a particular website is legitimate or fake. Alternatively, it also checks the authenticity of images and extracts the text present on that image and determines the text and classify if it is real or fake. Our system is also detecting the text information or content related to COVID-19 which is spread through social media. This system helps in reducing the circulation of false news, it prevents user from accessing fake duplicate website or URLs.

VI. FUTURE SCOPE

Recent development in technologies and the vast use of internet had a dominant effect in every sector of the economy. Unfortunately, technical advancements have been accompanied by modern advanced tactics for attacking and defrauding consumers. Here is where, our project comes into account.

- It will analyse the different URLs and provide the accurate result.
- System will avoid the use of fake duplicate websites.
- It would lessen the risk of phishing attack.
- It can detect misinformation and fake information.
- Anti-phishing browser can be developed using the same technique.
- More media files should be added to the system such as (Audio and Video Messages).
- Modern and upcoming future technologies can be used to improve the performance.
- Fact Checking should be improved using the third party APIs and can should make the system dynamic.

VII. REFERENCES

[1] An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features," Mohammad Mehdi Yadollahi, Farzaneh Shoeleh, Elham Serkani, Afsaneh Madani, and Hossein Gharaee; 2019 5th International Conference on Web Research (ICWR)-IEEE 2019

[2] Ankit Kumar Jain, B. B. Gupta; "A machine learning based approach for phishing detection using hyperlinks information" Springer International Publishing Group, 2018

[3] Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. PhishNet: Predictive Blacklisting to detect Phishing Attacks. IEEE INFOCOM.

[4] K. ARVIND, S. GOVARTHAN, S. KISHORE KUMAR, M. NAVEEN KUMAR, R.LAKSHMI; "FAKE NEWS DETECTION AND RUMOUR SOURCE IDENTIFICATION" International Research Journal of Engineering and Technology (IRJET) Volume: 06 Issue: 03 | Mar 2019

[5] Bireswar Banik and Abhijit Sarma; "Phishing URL detection system based on URL features using SVM" International Journal of Electronics and Applied Research (IJEAR) vol. 5, issue 2, Dec 2018

[6] Zaitul Iradah Mahid, Selvakumar Manickam, Shankar Karuppayah; "Fake News on Social Media: Brief Review on Detection Techniques"IEEE-2018

[7] Atik Mahabub, Springer-2020, "Using Ensemble Voting Classifier and comparison with other classifiers, a robust technique for detecting false news."

[8] Muhammed Afsal Villan, Kuncheria Kuruvilla, Johns Paul; "Fake Image Detection Using Machine Learning" International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 - 2017

[9] Aliya Begum and Srinivasu Badugu; "A Study of Malicious URL Detection Using Machine Learning and Heuristic Approaches" Springer ICETE -2019

[10] Kuai Xu, Feng Wang, Haiyan Wang, and Bo Yang; "Detecting Fake News Over Online Social Media through Domain Reputations and Content Understanding," IEEE-2020,TSINGHUA SCIENCE AND TECHNOLOGY

[11] Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh; "Detecting Phishing Websites Using Machine Learning" IEEE-2019

[12] Aditya Ambre, Praful Gaikwad, Kaustubh Pawar, Vijaykumar Patil."Web and Android Application for Comparison of E-Commerce Products", International Journal of Advanced Engineering, Management and Science(ISSN: 2454-1311),vol.5,no. 4, pp.266-268,2019.