Detection and Analysis of Fraudulent Content on Web Platforms

A thesis Submitted to

DELHI TECHNOLOGICAL UNIVERSITY

For the Award of degree of

DOCTOR OF PHILOSOPHY

In

DEPARTMENT OF INFORMATION AND TECHNOLOGY

by

DEEPIKA VARSHNEY

(2K18/PhD/IT/501)

Under the Supervision of

Prof. Dinesh Kumar Vishwakarma

Department of Information and Technology
Delhi Technological University
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042, India
DECEMBER 2021

DECLARATION

I declare that the research work reported in the thesis entitled **“Detection and Analysis of Fraudulent Content on Web Platforms”** for the award of the degree *of Doctor of Philosophy* in the *Department of Information and Technology* has been carried out by me under the supervision of Dr. *Dinesh Kumar Vishwakarma*, Professor in Department of Information and Technology, Delhi Technological University, Delhi, India.

The research work embodied in this thesis, except where otherwise indicated, is my original research. This thesis has not been submitted earlier in part or full to any other University or Institute for the award of any degree or diploma. This thesis does not contain other person’s data, graphs, or other information unless specifically acknowledged.

Deepika

Date: 9-01-2022

Deepika Varshney
2K18/Ph.D./IT/501

CERTIFICATE

This is to certify that the work contained in the thesis entitled **"Detection and Analysis of Fraudulent Content on Web Platforms"** submitted by Ms. Deepika Varshney (Reg. No.: 2K18/Ph.D./IT/501) for the award of degree of Doctor of Philosophy to the Delhi Technological University is based on the original research work carried out by her. He has worked under my supervision and has fulfilled the requirements as per the requisite standard for the submission of the thesis. It is further certified that the work embodied in this thesis has neither partially nor fully submitted to any other university or institution for the award of any degree or diploma.

Prof. Dinesh Kumar Vishwakarma
Supervisor
Department of Information Technology
Delhi Technological University, Delhi

ACKNOWLEDGMENT

I am grateful to my supervisor, **Prof. Dinesh Kumar Vishwakarma,** and DRC Chairperson without whom this achievement would not have been realized. It was his valuable guidance and consistent encouragement all through my research period, which helped me to overcome the challenges that came in the way. I consider it as a great opportunity to do my doctoral program under his guidance and to learn from his research experience.

My sincere regards to **Prof. Jai Prakash Saini, Vice-Chancellor, Delhi Technological University** for providing a platform for pursuing my Ph.D. work. I express my gratitude to **Prof. Kapil Sharma**, Head, Department of IT, DTU for their kind support and for providing the necessary facilities to undertake this research. I would also like to thank the senior staff members and lab technicians for helping me as per my requirement during the research work.

I do not have words to thanks my fellow lab mates, friends, hostel mates who offered me their time, help, and moral support whenever needed. I am grateful to my husband (Mr. Mohit Aggarwal), parents (Mr. Krishna Kumar Varshney and Mrs. Chandrakanta Varshney), and family for being an inspiration and strong support for me always and enlightening my path of life.

Last but not least, I am thankful to God (Lord Shiv and Krishna) for always giving me the strong support, the best path, and direction to make things successful.

Deepika

Deepika Varshney

ABSTRACT

Recently, the false information detection accompanying multimedia content entices numerous real-life applications such as election, emergencies, health care, terrorism, etc. One of the ultimate aims of artificial intelligence society is to develop an automatic system that can be recognized and understand fraudulent content accurately. Over the decade, many efforts made to recognize the false information accompanying multimedia data but still it is a challenging task to detect as sometimes sufficient evidence are not available to verify the content. To start with, we have reviewed the most popular and prominent state-of-the-art solutions, compared, and presented. Based on the literature survey, these solutions are categorized into handcrafted features-based descriptors and automatically learned features based on deep architectures. In this thesis work, the fraudulent content detection framework is divided into traditional machine learning(TML) and deep learning (DL) based architectures which are then utilized throughout this work.

The first chapter detailed discussed about the technique employed for the prediction of fraudulent content having text as an input. An overview of the complete model is described in the following paragraph. The techniques we covered here in this concern are based on two ways. In the first case, the input is given as text embedded images, while in the other case, in the simple text format. In the text embedded images, using the OCR technique (optical character recognition) the text content is retrieved from an image. Whereas, the second case considered the text only content. These two cases have been considered in this chapter and techniques involved in each of these cases have been discussed in detail. In the next chapter, we considered the claim accompanying multimedia content (images and videos). Here, firstly we discussed the technique where the claim accompanies image content, and secondly, the technique concerning to the claim accompanies video content. The third chapter, elaborates the proposed multi-web platform framework for detecting deceptive claims on the social media platform. Spreading of misleading information on social web platforms has fuelled huge panic and confusion among the public regarding the Corona disease, the detection is of paramount importance. Previous studies mainly relied on a specific web platform to collect crucial evidence for the prediction of misleading information. The analysis identifies that retrieving clues from two or more different web platforms gives more reliable prediction and confidence concerning a specific claim. This study proposed a novel multi-web platform voting framework that incorporates the 4 sets of novel features (including content features, linguistic features, similarity features, and sentiments features). To validate the claim, a unique source platform is designed to collect relevant headlines viz. YouTube and Google based on specific queries. The features are extracted concerning each collected headline. This unique platform can also help researchers to gather efficient headlines from various web platforms. After evaluation, it has been observed that our proposed intelligent strategy gives promising results and is quite effective in predicting misleading information. The model correctly detected about 98% of the COVID misinformation on the constraint Covid-19 fake news dataset. Furthermore, it is observed in our study that it is efficient to gather clues from multiple web platforms for more reliable predictions to validate the news. The proposed work provides practical implications for the policy-makers and health practitioners that could be useful in protecting the world from misleading information proliferation during this pandemic.

Finally, this thesis work is concluded with significant findings and future research aspects in the field of fraudulent content detection on social media.

List of Publications

International Journals/Conferences

1. Varshney, Deepika, and Dinesh Kumar Vishwakarma. "Hoax news-inspector: a real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles." *Journal of Ambient Intelligence and Humanized Computing* 12.9 (2021): 8961-8974. **Impact Factor: 7.104**
2. Vishwakarma, Dinesh Kumar, Deepika Varshney, and Ashima Yadav. "Detection and veracity analysis of fake news via scrapping and authenticating the web search." *Cognitive Systems Research* 58 (2019): 217-229. **Impact Factor: 3.52**
3. Varshney, Deepika, and Dinesh Kumar Vishwakarma. "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences." *Applied Intelligence* (2021): 1-22. **Impact Factor: 5.095**
4. Varshney, Deepika, and Dinesh Kumar Vishwakarma. "A review on rumour prediction and veracity assessment in online social network." *Expert Systems with Applications* (2020): 114208. **Impact Factor: 6.95**
5. Varshney, Deepika, and Dinesh Kumar Vishwakarma. "A Unified Approach of Detecting Misleading Images via Tracing its Instances on Web and Analysing its Past Context for the Verification of Content." *arXiv preprint arXiv:2109.09929* (2021). *(Communicated to International Journal of Multimedia Information Retrieval**, Impact factor: 3.250**, Under Minor Review)*
6. Varshney, Deepika, and Dinesh Kumar Vishwakarma. "An Automated Multi-Web Platform Voting Framework to Predict Misleading Information Proliferated during COVID-19 Outbreak using Ensemble Method ". *arXiv preprint* (2021)**.** *(Communicated to Neural Computing and Applications**, Impact factor: 5.606**, Under Major Review)*
7. Varshney, Deepika, and Dinesh Kumar Vishwakarma. "Analysing and Identifying Crucial Evidences for the prediction of False Information proliferated during COVID-19 Outbreak: A Case Study." *2021 8th International Conference on Smart Computing and Communications (ICSCC)*. IEEE, 2021.
8. Varshney, Deepika, and Dinesh Kumar Vishwakarma. "ArtiMarker: A Novel Artificially Inflated Video Marking and Characterization Method on YouTube." *2021 5th International*

Conference on Computer, Communication and Signal Processing (ICCCSP). IEEE, 2021.

List of Figures

List of Tables

Table of Content

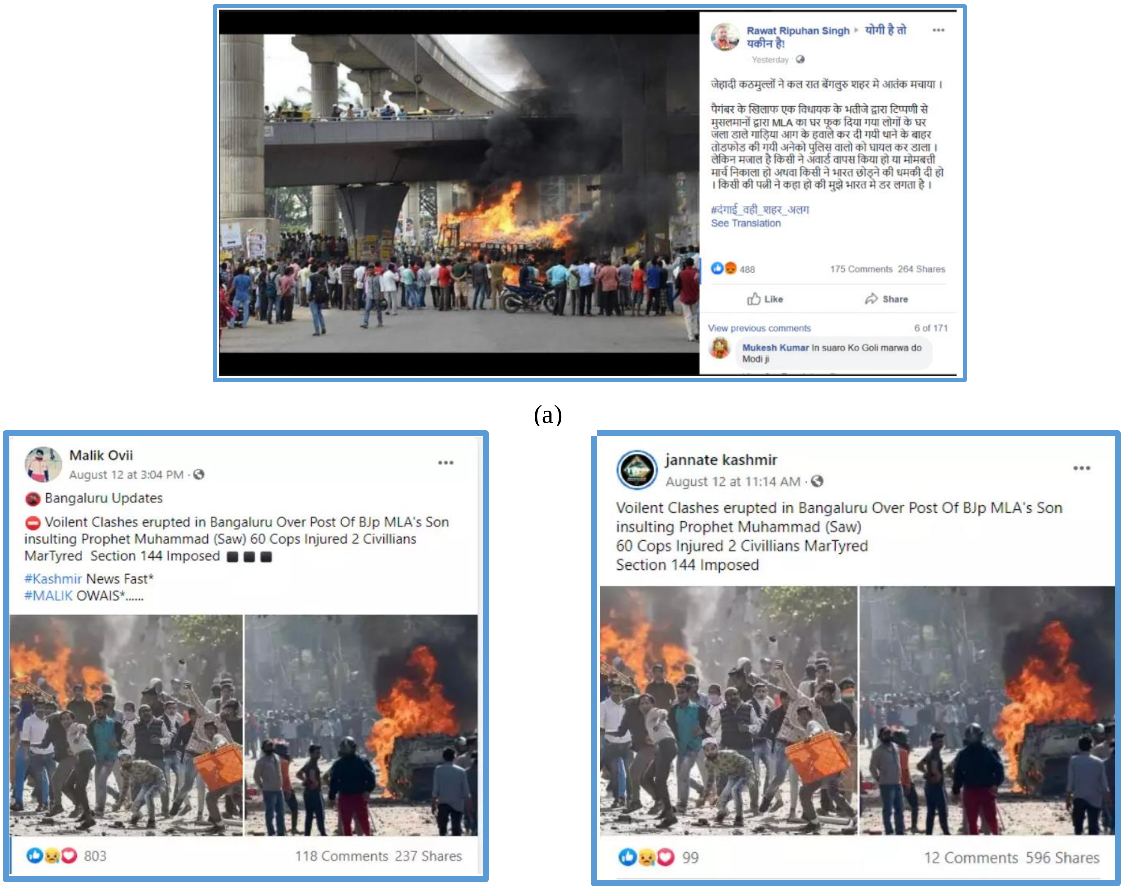
Chapter 1

Introduction

This chapter provides the fundamental concepts concerning fraudulent content detection. It involves basic terminologies, types, fundamental architecture, challenges, key contribution, motivation, application, and lastly, an overview of the remaining chapters. All these aspects are discussed in detail.

Online Social Networks (OSNs) play an important role in the way people communicate and consume information. This is mainly because OSNs provide an ideal environment for communication and information acquisition and users have access to a staggering amount of posts and articles that can share with others in real-time. To persuade users' decisions and opinions, online content has played a significant role in recent times [1]. Nowadays most people spent their time on OSNs to communicate with the world and use social media to engross news and seek out information instead of traditional news media. This is because the information propagation on social media takes very less time and is also less expensive rather than traditional news media, e.g, television and newspaper [2]. For instance, in the USA 49% of the adult population has accessed social media to share information in 2012, while in 2016 over 62% reported grasp news on social media on a daily basis [3]. Unfortunately, OSNs have also become the mechanism for massive campaigns to diffuse false information. During the last year, the rapid diffusion of false information raises serious concern because of the fact that social media plays a vital role to influence people's daily decisions in political, social, and economic domains. Therefore, false information detection in social media is a matter of concern [4]. It has been found that social media platform gives immense features and allow a user to share their thoughts and opinion with others in an easy way. For sharing any kind of information, people can have the following possible mindset. The first possible case is when users have something in mind and like to share the facts with their friends and group circle. The intention of sharing can be in a positive or negative way. In a positive way that information can be useful to the public, but if the intention was wrong than it can be used as a stepping stone to propagate false information as it create chaos and confusion among the public regarding an event. The false information may hamper public emotion, also create misconception among public. It can be dangerous as well as pathetic sometimes that can lead to be a reason of big war. One of the recent example is bangalore riots, where one of the old image on Facebook gone viral, that a truck set on fire in board daylight with the claim that it is from tuesday’s bengaluru riots, but actually the image is from past event (*2016 cauvery riots in Bengaluru*), reflecting in context to the current event as shown in Fig.1.1(a).

.



Rawat Ripuhan Singh ▸ योगी है तो यकीन है!

Yesterday

जेहादी कठमुल्लों ने कल रात बेंगलुरु शहर मे आतंक मचाया ।

पैगंबर के खिलाफ एक विधायक के भतीजे द्वारा टिप्पणी से मुसलमानों द्वारा MLA का घर फूक दिया गया लोगों के घर जला डाले गाड़िया आग के हवाले कर दी गयी थाने के बाहर तोडफोड की गयी अनेको पुलिस वालो को घायल कर डाला । लेकिन मजाल है किसी ने अवार्ड वापस किया हो या मोमबत्ती मार्च निकाला हो अथवा किसी ने भारत छोड़ने की धमकी दी हो । किसी की पत्नी ने कहा हो की मुझे भारत मे डर लगता है ।

#दंगाई_वही_शहर_अलग

See Translation

488 175 Comments 264 Shares

Like Share

View previous comments 6 of 171

Mukesh Kumar In suaro Ko Goli marwa do Modi ji

(a)

Malik Ovii

August 12 at 3:04 PM ·

Bangaluru Updates

Voilent Clashes erupted in Bangaluru Over Post Of BJp MLA's Son insulting Prophet Muhammad (Saw) 60 Cops Injured 2 Civillians MarTyred Section 144 Imposed

#Kashmir News Fast*

#MALIK OWAIS*......

803 118 Comments 237 Shares

(b)

jannate kashmir

August 12 at 11:14 AM ·

Voilent Clashes erupted in Bangaluru Over Post Of BJp MLA's Son insulting Prophet Muhammad (Saw)

60 Cops Injured 2 Civillians MarTyred

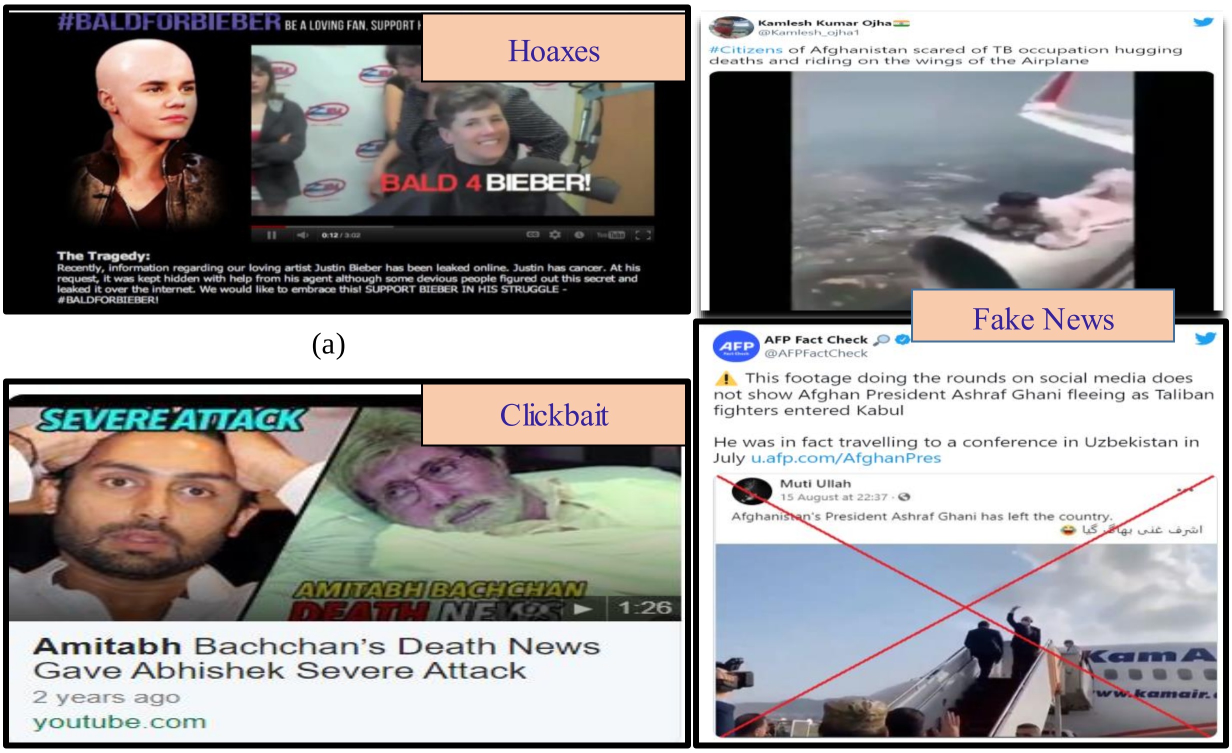
Section 144 Imposed

99 12 Comments 596 Shares

(c)

Fig.1.1. Example of Information Warfare, Fig.1.1(a) shows one of the old picture of 2016 Bangalore riots reflecting in context to the current event, Fig.1.1(b) and Fig 1.1(c) shows the riots that rocked Delhi is misrepresented as Bengaluru riots.

The misleading information creates lots of chaos and confusion in the mindset of people and is a reason for big war. Similarly, the pictures from riots that rocked Delhi in February 2020 is misrepresented as Bengaluru riots as shown in Fig.1.1(b) and Fig.1.1(c). In contrast to this, people may post on social media in concern to share some facts about a certain event with no negative intention. Whereas, some posts may share by users to clarify the facts regarding some event he/she is in doubt. The other cases can be possible when the user is in a doubtful stage and not confident about the facts presented. This type of information may not be intentionally spread, but disseminated because of the curiosity gap among the public regarding any news event. To validate the facts, the post is shared by a large number of people and if the information is not true then it could give a negative impact on society in many different ways. The false information can hurt the image of a candidate, altering the outcome of an election, in any terrorist attack and pandemic misleading information can cause widespread panic and general chaos. Following the same, some of the posts have been shared by malicious users intentionally to perform deceptive activities such as Rumors, Fake News, Clickbait's, Hoaxes, etc. The examples concerning the prominent forms of fraudulent content is shown in Fig.1.2. The Fig 1.2(a) shows an example of one of the Hoax goes viral that the popular star Justin Bieber was diagnosed with cancer and the #Baldforbeiber , goes popular that request people to go bald to show their support. Many people go bald after this news is propagated, however, after some time it has been identified that the news is a big hoax and not true.



#BALDFORBIEBER BE A LOVING FAN, SUPPORT H
Hoaxes
BALD 4 BIEBER!
0:12 / 3:02
Fake News
Clickbait
SEVERE ATTACK
AMITABH BACHCHAN DEATH NEWS
1:26

The Tragedy:
Recently, information regarding our loving artist Justin Bieber has been leaked online. Justin has cancer. At his request, it was kept hidden with help from his agent although some devious people figured out this secret and leaked it over the internet. We would like to embrace this! SUPPORT BIEBER IN HIS STRUGGLE - #BALDFORBIEBER!

Kamlesh Kumar Ojha
@Kamlesh_ojha1
#Citizens of Afghanistan scared of TB occupation hugging deaths and riding on the wings of the Airplane

AFP Fact Check
@AFPFactCheck
⚠ This footage doing the rounds on social media does not show Afghan President Ashraf Ghani fleeing as Taliban fighters entered Kabul

He was in fact travelling to a conference in Uzbekistan in July u.afp.com/AfghanPres

Muti Ullah
15 August at 22:37 ·
Afghanistan's President Ashraf Ghani has left the country.
اشرف غنی بھاگ گیا 😂

Amitabh Bachchan's Death News Gave Abhishek Severe Attack
2 years ago
youtube.com

(a)

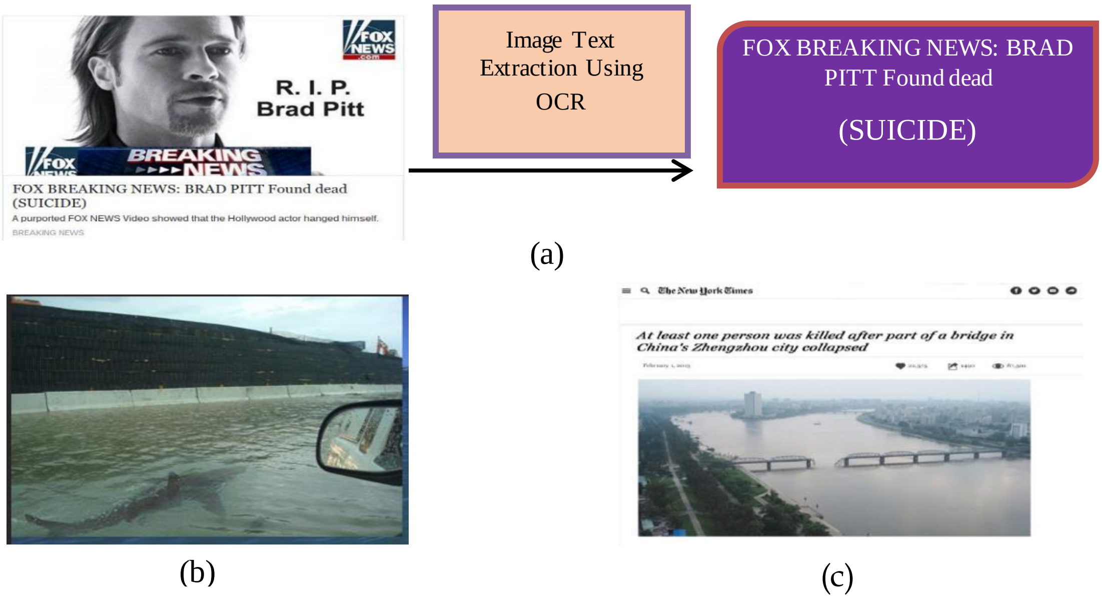
(c)

(b)

Fig.1.2. Example of different forms of fraudulent content, Fig.1.2(a) shows an example of one of the Hoax that justin beiber is diagnosed with cancer, Fig.1.2(b) false information propagated when the Taliban entered Afghanistan that citizens are scared of TB occupation and and false information about president Ashraf Ghani has left the country due to fear of Taliban. Fig.1.2(c) shows the wrong thumbnails that Amitabh Bachchan is dead.

Whereas, Fig.1.2(b) shows one of the recent examples of false information propagated when the Taliban entered Afghanistan that citizens are scared of TB occupation that hamper public emotions and false information about president Ashraf Ghani has left the country due to fear of Taliban. Similarly, the news is spread on social media sites with wrong thumbnails that Amitabh Bachchan is dead. These catchy headlines make users curious to view a video and malicious users use it as a stepping stone to increase their channel views as shown in Fig 1.2(c).

The malicious users intentionally weaponize information to do a diverse set of fraudulent activities on web platforms and have dire consequences to the public: mutating their opinions and actions, especially concerning critical world events like major elections. Therefore, false information detection in social media is a matter of concern. To share any information regarding an event there are many different ways in which people may share their point of expression either in support or not in support regarding an event by posting information in the form of text, text embedded with an image, or text attached with an image/audio/video, etc. The attached image may be tampered or photoshopped via some mechanism to reflect some false event as true. The Fig.1.3. shows an example of misleading images. The Fig 1.3(a) shows an example of text embedded image, where the false text is attached with an image that Brad Pitt found dead, similarly misleading images also incorporates tampered/manipulated images. Automated software has been used to tamper a specific portion of an image and post fake reports on social media, the example shown in Fig.1.3(b) and Fig.1.3(c). The Fig.1.3(b) represents an example of the manipulated image, it shows a spliced shark on a photo during Hurricane Sandy in 2012, Fig.1.3(c) shows the fake news image that was modified to depict a bridge collapse. A detailed description of the definition, types, and fundamentals concerning fraudulent content as well as the basic framework of fraudulent content detection has been discussed in the detail in the following subsection.


R. I. P.
Brad Pitt
FOX NEWS
BREAKING NEWS
FOX BREAKING NEWS: BRAD PITT Found dead (SUICIDE)
A purported FOX NEWS Video showed that the Hollywood actor hanged himself.
BREAKING NEWS
Image Text Extraction Using OCR
FOX BREAKING NEWS: BRAD PITT Found dead
(SUICIDE)
The New York Times
At least one person was killed after part of a bridge in China's Zhengzhou city collapsed

(a)

(b)

(c)

Fig.1.3. Examples of misleading images, Fig.1.3(a) shows an example of text embedded image, Fig.1.3(b) represents an example of the manipulated image, it shows a spliced shark on a photo during Hurricane Sandy, Fig.1.3(c) shows the fake news image that was modified to depict a bridge collapse.

1.1 Fraudulent Activities: Definition Perspective

In the new era of the internet and technology where people are exchanging their thoughts and opinion via social media sites, the credibility of information is a major challenge. As various actors are weaponizing information to deceive public opinions by publishing any content that does not faithfully represent the event that it refers to. Fraudulent post can be defined as:

"***Any post that shares content that does not faithfully represent the event that it refers to***".

This could include:

- Content from a past event that is reposted as being captured in the context of a currently unfolding similar event.
- Content that is deliberately manipulated (also known as tampering, doctoring, or photo shopping).
- Content that is published together with a false claim about the depicted event.

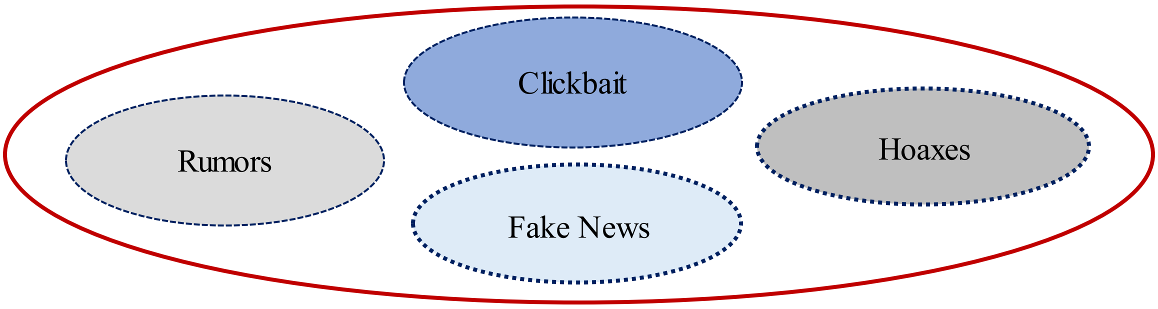


Fig.1.4 Forms of deceptive content

Different forms of fraudulent content have been identified in the previous studies. Some of the prominent forms of fraudulent content are Hoax, Clickbait, Fake News, and Rumor as shown in Fig 1.4.

Earlier studies have been provided definitions concerning each of these diverse sets of deceptive activities in different ways. The detailed definition w.r.t each of the forms of fraudulent content is shown in Table 1.1.

Table 1.1. The fundamental definition of the diverse set of deceptive activities

Category	Definitions
Fake News	- False or misleading content presented as news and communicated in formats spanning, written, printed or digital communication. - New stories that are fabricated that obtain little to no verifiable facts.

	- Stories that are probably false, have enormous traction in the culture, are consumed by millions of people. - Fake news is a news article that is intentionally and verifiably false and could mislead readers. - Fake news is created with malicious and dishonest intentions to mislead consumers.
Rumor	- Rumors are the unverified and doubtful information that passes on among the public without the intention to mislead. - a source that traffic in rumors, gossip, and unverified claims. The claim whose veracity is not yet cleared, and it is unknown. - In the social sciences, rumors can be defined as a form of a statement whose veracity is not quickly or ever confirmed. - A piece of unverified information of uncertain origin usually spread by word of mouth. - information or opinion that is widely disseminated without any authority or confirmation of accuracy.
Clickbait	- It is the attractive/catchy headline, the post with misleading thumbnails, questionable headlines that breaks the curiosity gap. - Misleading thumbnail link that is designed to attract attention and to entice users to follow that link and read view or listen to the linked piece of online content. - Clickbait's are the ridiculous, and misleading headlines that trick you into opening an image, video or, article link. - Clickbait's are the catchy headlines that are frequently used by social media outlets to lure its viewer into clicking them and thus leading them into dubious content. - It is a click technique in which a user manipulates the curiosity of a person in order to open more pages in a web site. - Clickbait is a bad habit of today's web publishers that resort to such a technique in order to deceive web visitors and increase publishers' page views and advertising revenue. - Clickbait is the term that is used to describe deceiving web content that uses ambiguity to prompt the user into clicking a link.
Hoax	- It is a falsehood deliberately fabricated to masquerade as the truth. - It is the news that contains false or inaccurate facts that are either inaccurate or false but which are presented as genuine. - It presents a half-truth used deliberately to mislead the public. - It can be anything that would elicit fear, make you angry, or seems important and that make you forward, reply, or action without first validating the information sources. - Hoaxes are non-malicious viruses whose main intent is to deceive human perception by conveying false information as truth. - Hoaxes are the unsolicited or unwanted emails which directly/ indirectly sent by personnel, they are the smarter version of spam that masquerade themselves well via the personnel that is present as ones' contacts. - A hoax can be defined as a try to convince any readers to believe particular deception.

1.2 Fraudulent Activities: Comparative Study

There are a diverse set of deceptive activities on social media that we have also discussed in the earlier section. These activities may differ from each other in some aspects like aim, intention, etc. In this section, the comparative study among the prominent form of fraudulent content such as fake news, clickbait, rumor, satire, and hoax are presented in Table 1.2.

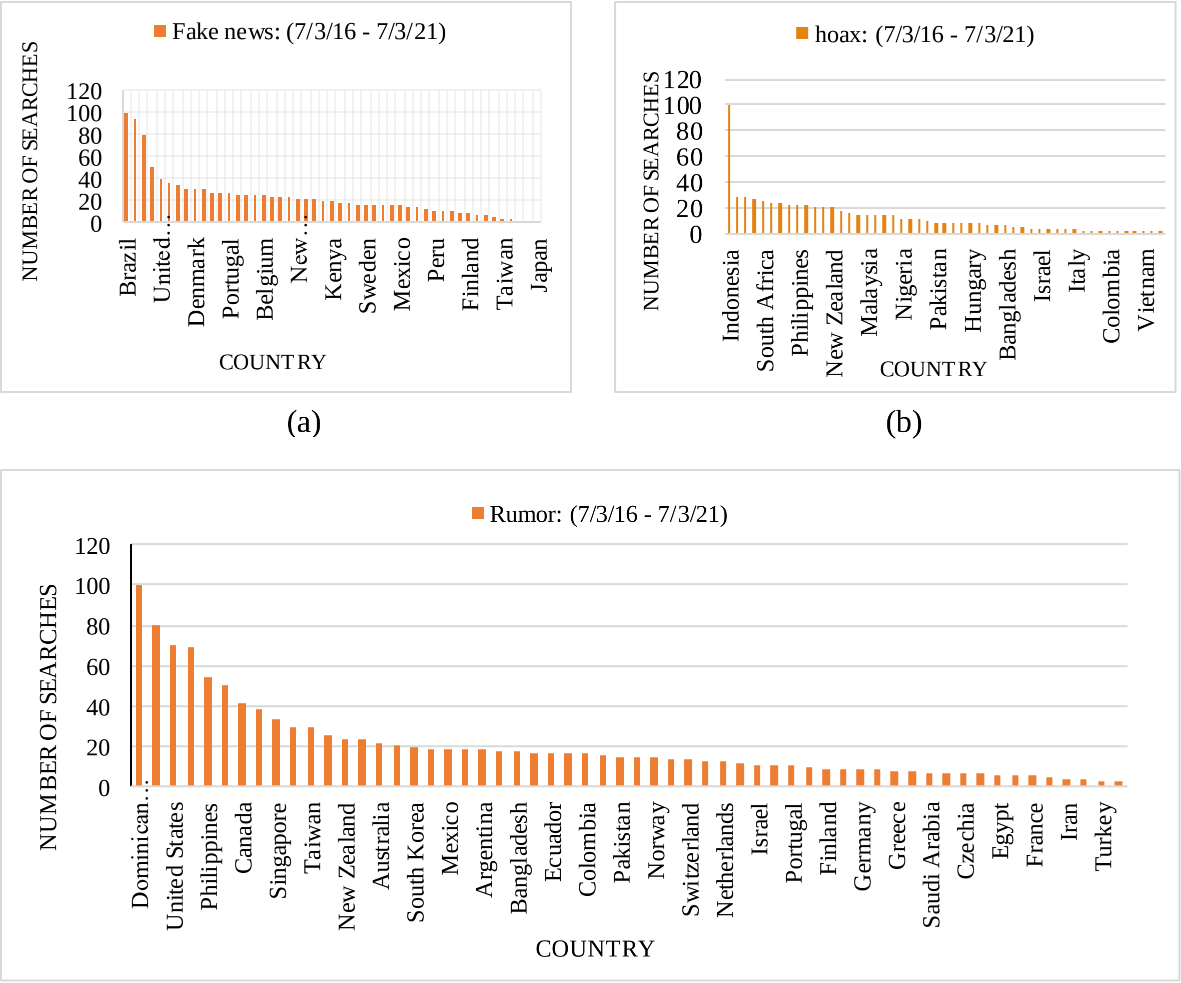
The deceptive activities are interrelated but differ w.r.t their aim and intentions. It has been identified that intention and purpose play a major role in differentiating one from another. Like, from the study it has been found that fake news, clickbaits, and hoaxes are intentionally spread and people do it with some bad intentions, whereas rumor and satire are unintentional, news is spreaded without verification. In the thesis, we have considered the problem as binary class classification and it can be resolved in two ways: either as true (factual)/ false (non-factual).

Table 1.2. Distinguishing different forms of fraudulent content

Category	Intentional?	Aim	Actors Involved
Fake News	YES	- Damaging the reputation of a person or entity. - making money through advertising revenue. - Political agendas and communal hate-mongering to scamming.	- Propagated by hostile foreign actors particularly during elections. - Spread by a people who believe them to be true. (Rumor) - - Spread by people who know it to be false. (Fake News) - Bots - Criminal/Terrorist Organization - Activist or Political Organization - Government - Journalists - Useful Idiots
Rumor	NO	- Intimidate others - Gain status - Seek attention from crowd - Part of political propaganda - Manipulate the victim/situation.	
Clickbaits	YES	- The teaser aim is to exploit the “curiosity gap” providing that much information that make readers curious, however the news is not enough to satisfy their curiosity without clicking through the linked content and may used for phishing attacks. - To gain monetary benefits.	
Satire	NO	- Fun - Main purpose is often constructive social criticism, using wit to draw attention to wider issues in a society and having an intent of shaming individuals, corporations,	
Hoax	YES	- The main goal of spreading hoaxes is to injure an organization, individual, or person - Financial or political benefit to maximize readership.	

1.3 Fraudulent Activities: Statistical Analysis

From the Google trend analysis, it has been observed that there is an immense search for a diverse set of fraudulent content w.r.t different countries.



(a)

(b)

(c)

Fig.1.5. The number of hits for deceptive activities w.r.t each country, Fig.1.5. (a) Hits concerning to fake news, Fig.1.5. (b) Hits concerning to Hoax, Fig.1.5(c) hits concerning to rumor[1].

Each of these countries has different search interests concerning diverse forms of fraudulent content. The trend analysis for each set of fraudulent content from the period of 2016-2021 is shown in Fig 1.5. It can be observed from Fig.1.5, that the fake news term is more searched in Brazil in comparison to other countries[1], whereas Indonesia has more number of searches concerning hoax forms of content. On the other hand, Dominican and United States people show their interest more towards rumors in terms of searches.

[1] https://trends.google.com/trends/?geo=US.

From these statistics it can be clearly seen that the different form of fraudulent content is searched over web from different countries. This shows an immense interest of countries towards diverse set of deceptive activities. Whereas it has been observed that each of these forms of fraudulent content is interrelated, however, they differ w.r.t their aim and intention. As from the previous discussion we found that fraudulent activities affect society and widely propagate on various social media, the detection and correction of these is of paramount important and need to be addressed.

It can be observed from the past researches that many countries are working to address fraudulent content as well as different application of it. The contributed research statistics report from (2016 – 2021)[2]has been shown in Fig.1.6.

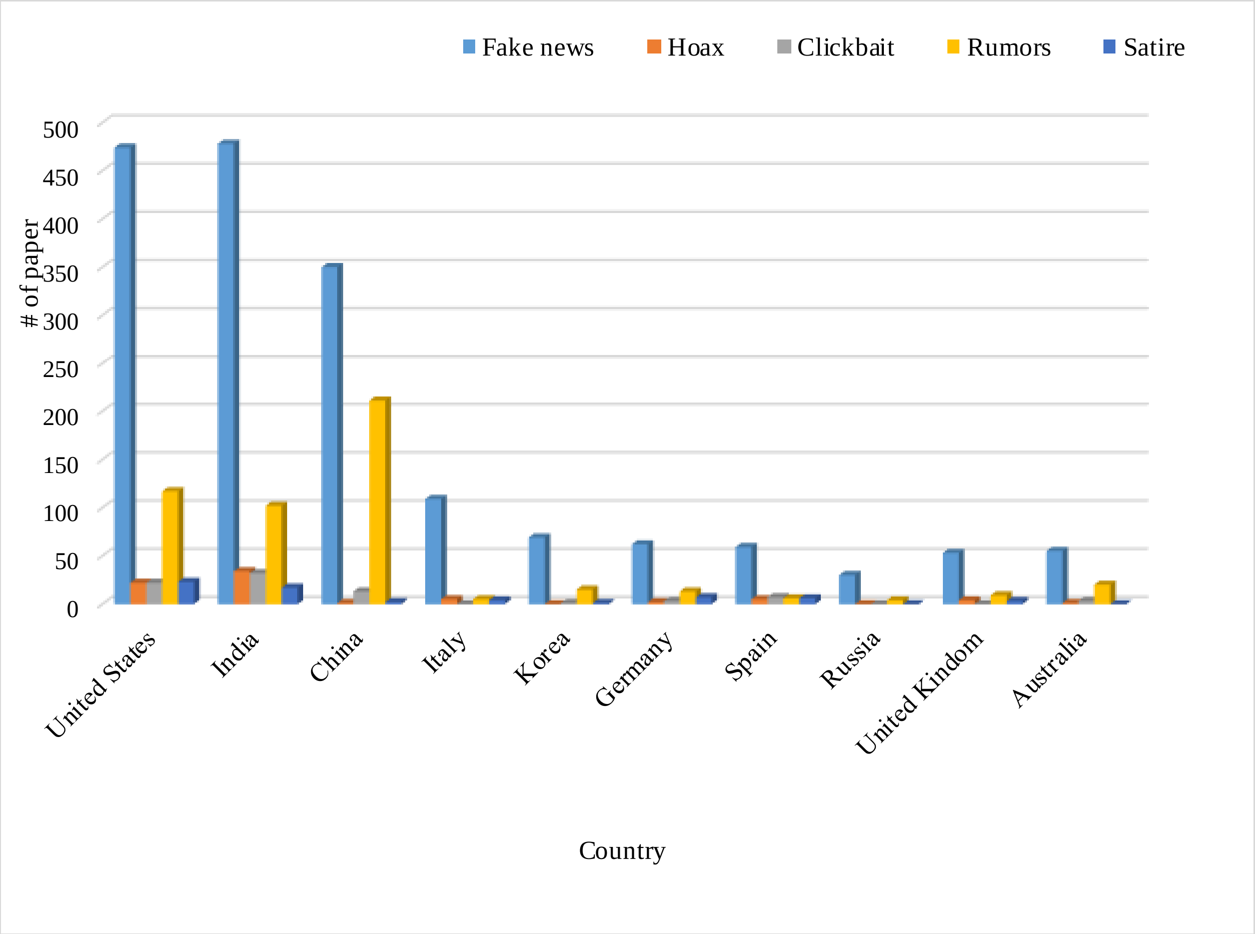


Fig.1.6. Statistics report of number of papers published (2016-2021)

[2] Engineering Village - Quick Search

1.4 Popularity of Web Platforms

Social media platform provides an easy way of sharing information among other groups and people. The platforms such as Facebook, Twitter, YouTube provide a way to users for expressing their point of expression with the public. However, the aim of sharing the thought maybe with some different intentions. Firstly, the user wants to share some information with no bad intention, the second possible case is when the information has been shared with some bad intention for personal benefits and lastly, it may be the case when the user share some information but doesn't know whether the information is true or not, the veracity of the news is not clear, however in this case the intention of the user may not be wrong. The social media platform has been widely utilized by users. The popularity of social networks worldwide as of July 2021, ranked by the number of active users as shown in Fig.1.7, where it can be seen that Facebook has the highest number of active users in million among other platforms and widely used social media platforms for sharing content. From the analysis, it has been observed that the total population is 7.799 billion(approx.), 4.66 billion active internet users, 4.2 billion active social media users, and 4.32 active mobile internet users[3] as shown in Fig.1.8. There are various social media platforms available that allow the user to share information, like some of the examples are Facebook, YouTube, etc., which are prominently used worldwide by the users[4].

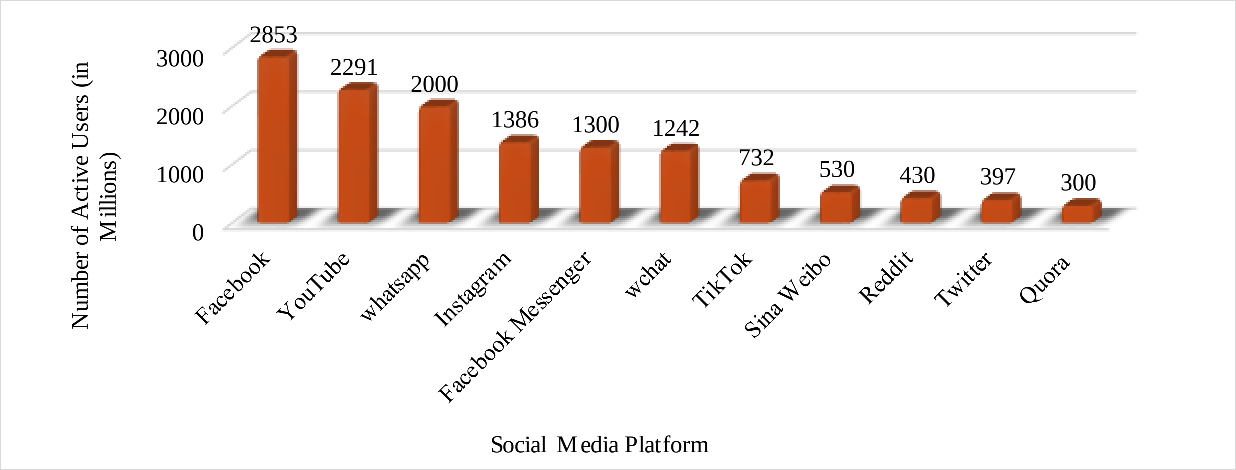


Fig.1.7. Number of active users in million

[3] https://www.internetlivestats.com/

[4] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

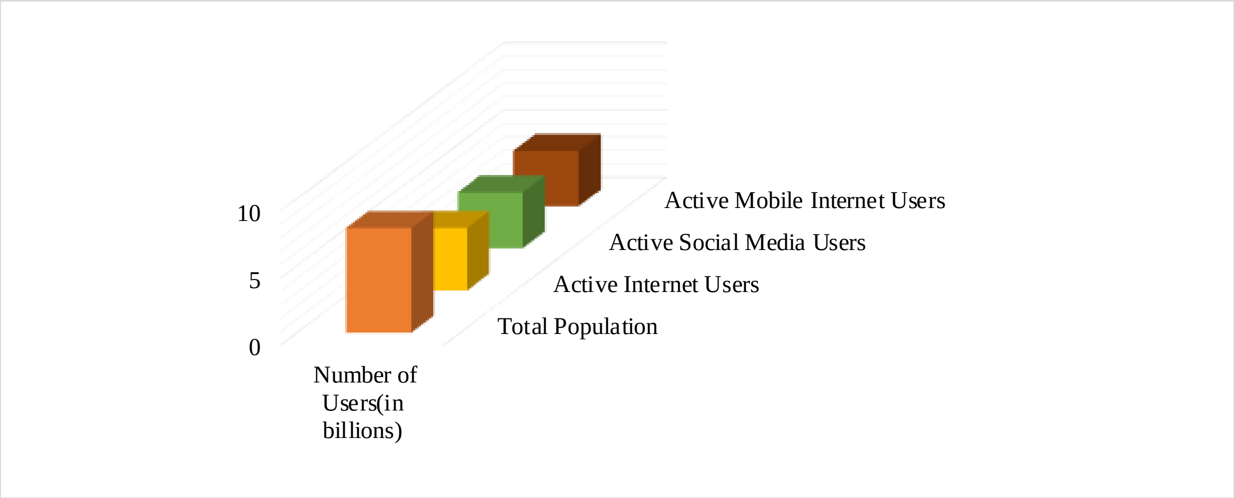


Fig 1.8. Statistics of Internet Users

1.5 Basic Fundamental Framework for Fraudulent Content Detection

The basic fundamental framework for FC detection has been discussed in Fig.1.9. In online social networks, the data can be available in many different modalities either in the form of text, text embedded images, images, videos, etc. Each of these modalities of data has been processed in different ways. In the thesis, we have considered text data as an input and different modalities of data have to be converted into text format (if it is in the form of text embedded images then the text has been extracted from an image using optical character recognition technique. If it is a video content, then google speech to text API (application programming interface) has been used to convert speech into text). There can be two frameworks that can be applied here either a traditional machine learning(ML) framework or deep learning(DL) framework for FC detection. In the machine learning paradigm, the text is first going to pre-processing phase, where the redundant and unnecessary words are removed like stop words, and other techniques are also applied such as lemmatization, stemming, etc. In the second part, the processed text is going to the feature extraction phase where the crucial evidence has been retrieved and further the retrieved features are utilized to train the model for the classification of data into credible and misleading. The testing has been done to evaluate the performance of the model.

On the other hand, in DL framework, the text is passed to the embedding layer, where the words are converted into vectors which further goes to the hidden layer to extract the complex hidden patterns from the input data. Lastly, the output is fed to a dense layer to classify the input as fake or real.

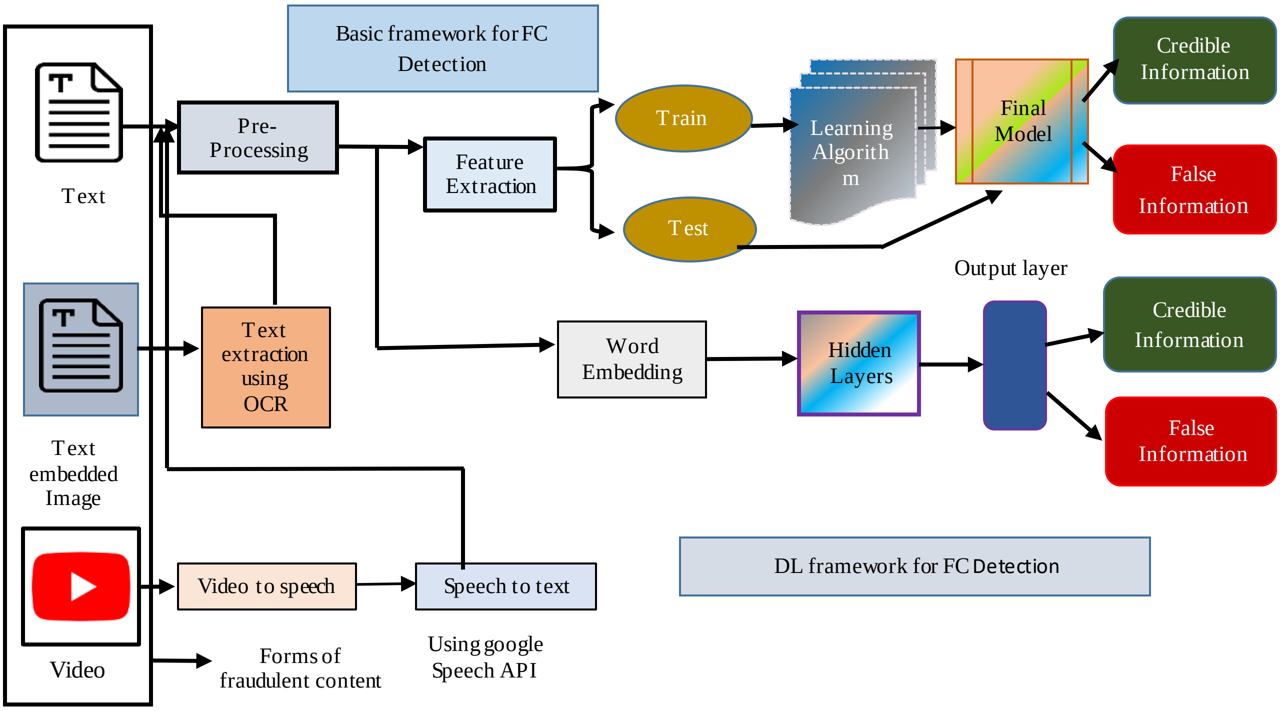


Fig.1.9. Basic Fundamental Framework for FC detection

1.6 Approaches and Contribution

The thesis incorporates the models for the detection of fraudulent content in the online social network. In the current scenario, people share their expression of thoughts in many different forms either in the form of text, claim accompanying with text/audio/video. The multimedia content along with the claim, attract users to any news article as image/ videos/ audio are a more convenient way to represent any event. The different machine learning and deep learning methods have been applied, incorporating a diverse set of features including content, semantics, linguistics, user profile, video content, etc. The publicly available datasets are identified and their parameter settings as well as challenges have been tabled. In this thesis, the work is contributed in three different modalities of data. In the first case the event is represented only w.r.t the text, whereas the second case considered the claim accompanying image on web platforms. The third case incorporates the claim accompanying video content. Whereas, the fourth case explores the text-embedded images. Along with this, the analysis has been done by incorporating a multi-web platform framework, where the evidence has been gathered/extracted from more than one web platform for the detection of fraudulent content. The various standard datasets have been used to validate the performance of the model. A detailed description of the proposed methodologies has been discussed in the later section.

1.7 Application of Detecting Fraudulent Content

The detection and analysis of fraudulent content have a wide range of applications. Some of the prominent applications where it can be applied i.e. in the election, Terrorism, Emergency Situation, Health Emergency, etc. As we have seen in the recent example of COVID 19 that lots of audio and video messages are propagated to mislead people regarding how to get rid of the coronavirus. One of the audio clips gone viral on WhatsApp and widely shared on Twitter and YouTube, that attributed to Dr. Devi Shetty Chairman and founder of Narayan Health, advises everyone "who has the coronavirus or is suspected of it should not go to get tested[5]", which later turned out to be false. This creates lots of misconceptions among the public. The normal user cannot able to verify the news, as there is no real-time tool or extension are available for analyzing multimodal data. However, some tools are available to process text[6]. There are a variety of applications where there is a need to implement an efficient framework for FC detection, some of the crucial areas are the election, healthcare, natural disaster, terrorism, etc. A recent example of COVID-19 pandemic, where lots of health-related rumors are spreading, pretending to be posted by some government officials[7]. In this situation, where people have an eye over any news announcement related to corona, malicious users use this opportunity to mislead people. One of the recent news reports that malicious users are sending emails purporting to be from HR departments, executives, and health organizations[8] and playing with human psychology to make them believe that the mail is coming from some government organization. These areas are still open to research.

1.8 Motivation of Detecting Fraudulent Content

The study of detecting fraudulent content on multimedia data is one of the important areas to study. The key motivation for detecting false information is as follows:

- To reduce controversies that have been created among people due to wrong information.
- Sometimes content is posted to lure the user to open the article but the content doesn't match with the heading information and degrades user experiences for ex. Clickbaits.
- False information sometimes emotionally hampers the public[5].

[5] https://www.altnews.in/fake-audio-clip-attributed-to-dr-devi-shetty-advises-against-getting-tested-for-coronavirus/

[6] http://twitdigest.iiitd.edu.in/TweetCred/

[7] https://www.buzzfeednews.com/article/janelytvynenko/ftc-fda-scam-coronavirus-cures

[8] https://www.buzzfeednews.com/article/janelytvynenko/coronavirus-fake-news-disinformation-rumors-hoaxes

- To maintain the reliability of the content on social media, as every person want authentic data. Due to the diffusion of false information, social media platform loses the trust of users.
- Due to the wide exposure of social sites such as Twitter, Sina Weibo, Facebook, etc., blogs, microblogs, and opinion sharing websites that allow people to freely express their thoughts, emotions, and feelings regarding any event. The exploration and diffusion of information across social media sites lead to gathering more attention towards research. As social media platforms are widely utilized for spreading breaking news and trending conversations that may contain unverified content related to incidents that happened in the real world.
- The emergence in the interconnection of the networks leads to high risk [6],[7],[8] of danger such as rumors, viruses giving a bad impact on society. The earthquake in Chile in 2010 is one of the examples, where the dissemination of rumors over social sites created a lot of chaos and confusion among people.
- The latest incident of misinformation that took place in Chicago on the second weekend of August 2018 related to 600 murders gives a serious destructive and negative impact [9] over society and presents fear and anxiety among people.
- Nowadays, people often look for information and knowledge related to health care from online social media, but not all of these sources provide reliable information that leads to the propagation of rumors.
- It can be seen from the Fig.1.10, that in 2015 to 2020 people are more aware of the rumor compare to fake news, however, after U.S presidential election in 2016, there are subsequent controversies that attracted an enormous interest importantly towards fake news in the last two years. Fig.1.11. shows another analysis over the past 12 years that depicts that research is steadily increasing in this area, which motivates our work.
- The online sharing of information among users generates a massive volume of data on social networks. This allows users to post their thoughts and opinions directly without any trusted external control due to which, it is one of the stepping stones to encourage misinformation.

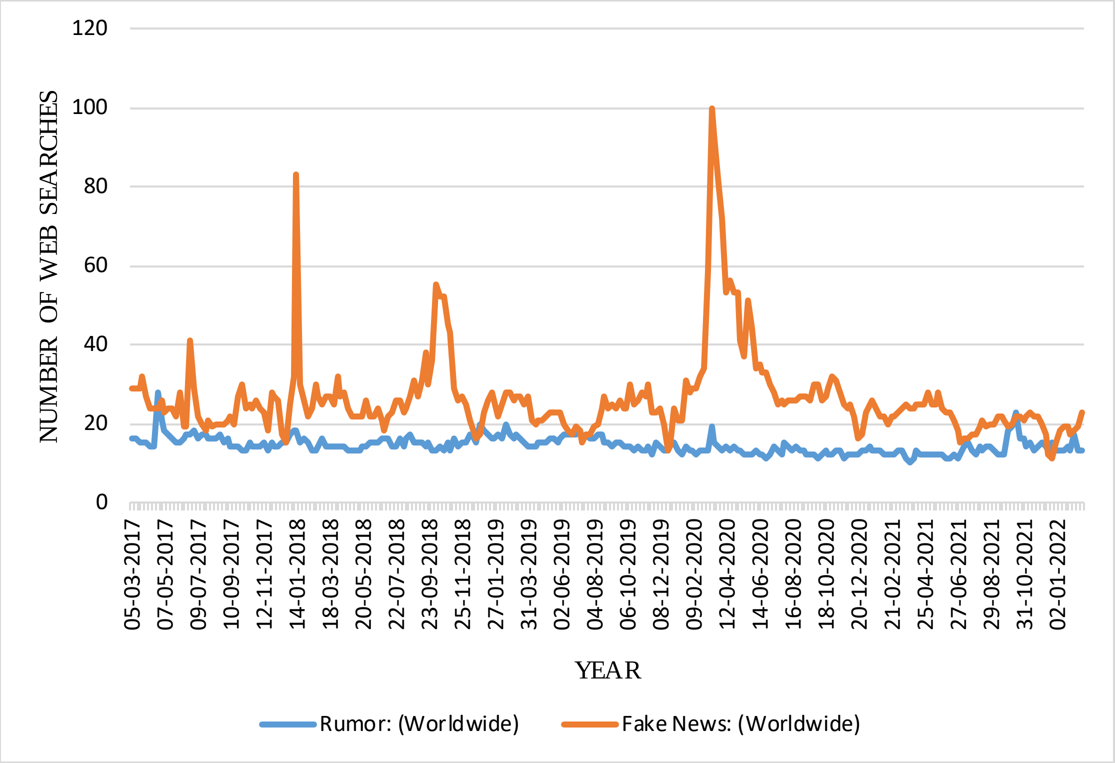


Fig.1.10. The number of search on rumor and fake news by the user in past years

From the above points, it can be observed that detecting rumors at an early stage is one of the crucial tasks. To reduce the spreading of misleading multimedia content, the online social media platform can adopt the following proposal.

- In the first proposal, the social media platform should apply the policy of not publishing any content without verification from the authentic source. The given proposal is good to counter fake but very less flexible, as now people are not being able to freely express their thoughts, which is one of the important aims of the social network. Hence, it may refuse by the social media platform.
- In the second proposal, the social media platforms should adopt the policy to first apply check over the content, before publishing. The posted content is not harmful or contains false information.

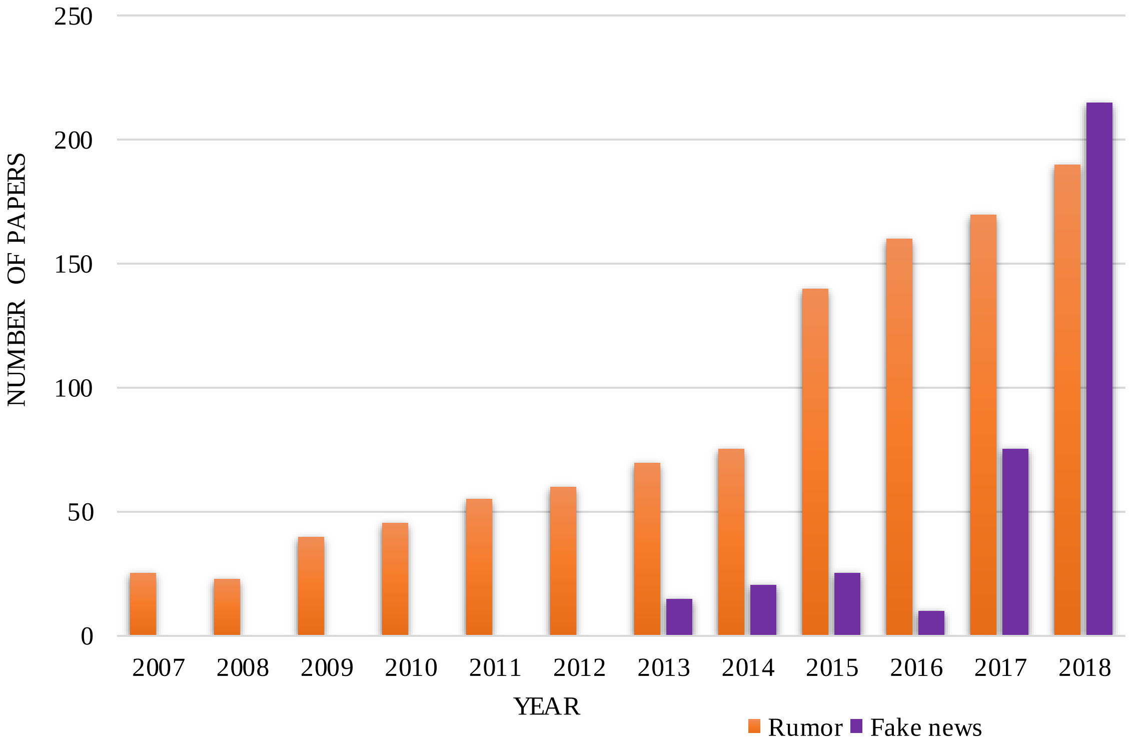


Fig.1.11. Progress of research over the past 12 years in the area of rumors and fake news.

1.9 Challenges associated with the credibility assessment of the information over the social web

In the era of the internet and technology, where users have adopted social media platform as one of the good sources of information and give an open opportunity to the influencer for spreading false information over the social media. This creates crises and other social problems that make FC detection over the social network as one of the important issues to address. Some of the challenges[1] associated with the credibility assessment and detecting truthfulness of the information over the social web are:

- Due to the complex network, it is quite difficult to identify resources useful in studying credibility.
- Various factors-like user behavior, preferences, and context that continuously influence the user's credibility.
- Lots of malicious and spam activities are going on over social networking platforms, which leads to inflate user popularity using some automated software or via using third party service.

- One of the challenges in the identification of rumourous tweets collected from Twitter is that it is difficult to characterize the content of a rumor in a way that it can be easily retrievable directly through an input query.
- Due to the problem in terms of the resources. Many of the researchers are restricted concerning the extent to which they can extend their work (OSN limitation). One of the crucial barriers to the development of a rumor analysis system is the lack of publicly available datasets.

The credibility assessment of information over the social web is one of the emerging fields; most of the researchers have shown their keen interest in it. According to [1], verification of information extracted from social media sites has become a very challenging task that needs to address.

1.10 Overview of Chapters

The rest of this document is organized as follows:

- Chapter 2 Literature review incorporates the existing state-of-the-art methods employed for the prediction of fraudulent content (TML and DL techniques).
- Chapter 3 describe an effective technique for detecting deceptive claim on web platforms and its evaluation in detail.
- Chapter 4 describe an effective technique for detecting deceptive claim accompanying multimedia content (Images and Videos) on web platforms and its evaluation in detail.
- Chapter 5 discusses the Multi-Web Platform framework for detecting deceptive claims.
- Chapter 6 concludes with future work.

Chapter 2

Literature Review

This chapter incorporates the existing state-of-the-art methods employed for the prediction of fraudulent content (TML and DL), existing available datasets (in the form of text, images, video), a diverse set of features used for the classification, research gaps and finally we end it with our research objectives.

Nowadays, detection of fraudulent content is one of the challenging tasks, as in the era of the internet and technology people are widely sharing the post with their group and sometimes people are sharing posts without verification that leads to making it viral in a few seconds. If the post is shared with some malicious intent, then it may hamper public emotions. There are a diverse set of deceptive activities propagating on social media, that differ from each other concerning their role. Some of the prominent forms are rumor, fake news, clickbait and a hoax. These concepts are interrelated with each other, however differ based on how these terms and concepts are defined. The study provided by [10], distinguish these terms and concepts based on three characteristics: (i) authenticity, (ii) intention, and (iii) whether the information is a news. Many researchers have put their keen interest towards the detection of fraudulent content by employing TML and DL techniques. From the previous literature, the following taxonomy has been designed as shown in Fig 2.1. The taxonomy is segregated concerning various aspects. The very first aspect is the types of false information and in this thesis, we primarily focus on some prominent forms of false information such as Rumour, Fake news, Hoax, and Clickbait. Whereas, the second aspect incorporates features employed for the detection of false information. A detailed description of the diverse set of features has been discussed in the later sections. To address the problem of fraudulent content, studies have been provided by employing Machine and Deep learning techniques. Concerning the TML techniques, the explored methods are: (1) Supervised; (2) Unsupervised; (3) Semi-supervised learning algorithm. From the previous studies, it has been found that Support Vector Machine, Naïve Bayes, K-Nearest neighbour, and logistic regression are some of the prominent machine learning techniques that have been applied in the identification of fraudulent content [11].

2.1 Data Collection Strategies

Social media platforms are prevalent nowadays as they provide an easy way for the user to express their personal opinion about any event, this can also be one of the stepping stones to spread false information and that's why collecting a good amount of data from social media platforms enables the research of detecting deceptive activities. Social media platform provides an easy and fast way to collect and store data via an APIs. However, the data collection is not only restricted via APIs. There can be many different ways of collecting data from social media and web platforms. Some of the prominent ways have been discussed in the following sections [12].

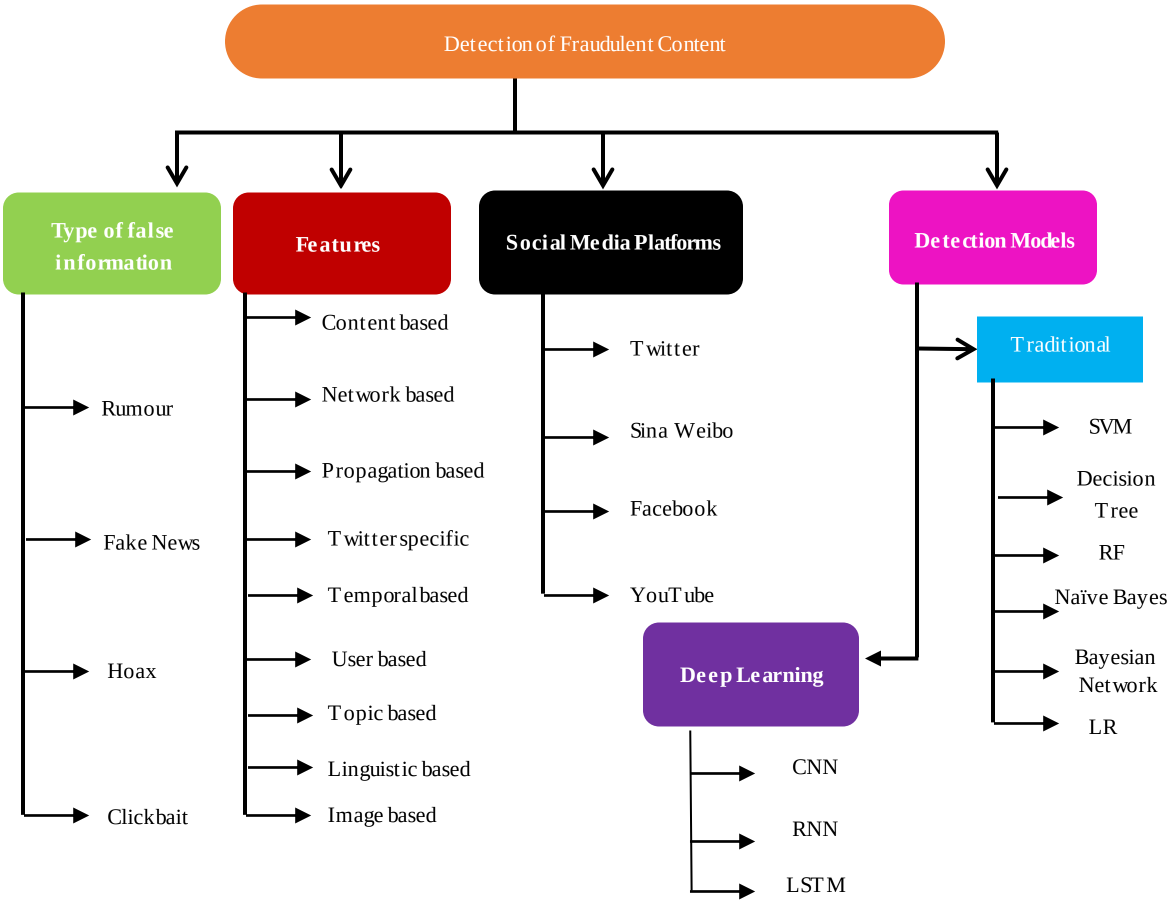


Fig 2.1 Taxonomy of Fraudulent Content Detection

2.1.1 Access to Social Media Data

There are many different ways of accessing data from social media websites, some of the prominent methods of accessing data are shown in Fig.2.2. One of the ways of accessing data is via APIs. Online Social media platforms such as Twitter, Sina Weibo, Facebook, etc. provide APIs that are designed to be accessed by other software applications. The well-defined methods are provided by an API using which the application can request data access. The complete documentation has been provided by each platform to let the user understand how to request data of interest like for Twitter[9], Sina Weibo[10], and Facebook[11]. False information spreading is a significant issue over the social media platform, and this is the key to utilize social media data for misleading data analysis. There are mainly three prominent key platforms (Facebook, Sina Weibo, and Twitter) in which study/analysis has been performed by the different state of the art methods [13]. From the Twitter report, it has been observed that nearly 95 million tweets per day have been published by its users, which makes Twitter one of the stepping stones to encourage misinformation[14]. Like Twitter, Sina Weibo is also one of the most popular Chinese microblogging platforms, however, has restrictions over some of its methods [15]. Moreover, Sina Weibo offers an officially rumor-busting service, which is not provided by many of the other social media platforms. The other ways of accessing the social media content are through scrapping the web (some of the prominent libraries in python are beautiful soap, scrappy, etc.) or via automatic testing tools like selenium web driver.

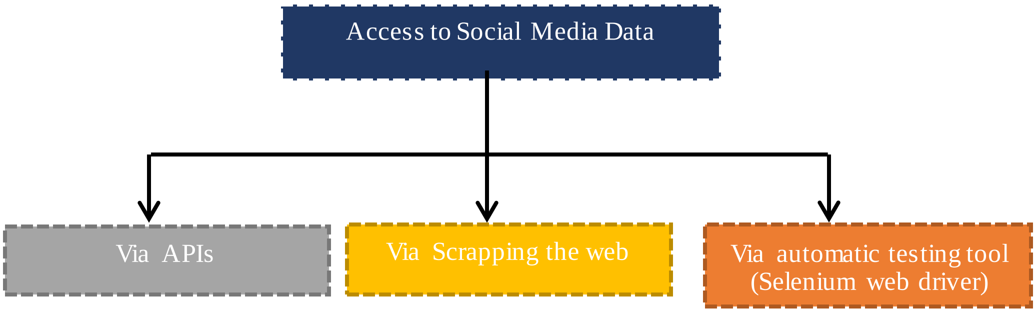

Access to Social Media Data
Via APIs
Via Scrapping the web
Via automatic testing tool (Selenium web driver)

Fig.2.2. Prominent ways of accessing the data from social media.

2.1.2 State-of-the-Art Data Collection Approaches

It has been noted that a careful data collection strategy is required to build up good datasets with relevant information for the development of an effective FC detection system. For

[9] https://dev.twitter.com/docs
[10] http://open.weibo.com/wiki/API%E6%96%87%E6%A1%A3/en.
[11] https://developers.facebook.com/docs/

experiment and evaluation mostly authors have used key social media platforms such as Twitter, Sina Weibo, YouTube, and Facebook or rumor debunking websites (Snopes, Politifact, Fact Check, etc.) for data collection[13], where many of them applied keyword-based filtering approach to collect data relevant to an event [16]. The detailed description of the data collection approach adopted via various state-of-the-art methods is shown in Table 2.1. Four crucial factors have been considered, concerning the data collection process. The first factor is the detection model that describes which learning algorithm is employed (TML, DL, or hybrid algorithm). The second-factor incorporates the type of platform has been exploited and the last factor describes the event considered for the data collection. It can be found that most of the previous studies have adopted Twitter platform for the data collection, due to easy access to data and everything is publicly accessible, whereas there is some restriction on other platforms. It is evident from the Table 2.1 that earlier TML is widely adopted learning mechanism, however, sometimes it is difficult to gather efficient hidden clues from the data via TML methods, due to which the research is also shifting towards deep learning approaches in the last 2 years, to get hidden representation from the text data and to extract more valuable information. Along with the collection of data schemes, some of the prominent existing datasets that are available for detecting deceptive activities on multimedia data is shown in Table 2.2, where it can be seen that there are very few standard datasets are available for fraudulent content detection, especially on images and requires further attention.

Table 2.1. Presents the factors concerning the data collection process adopted by existing state-of-the-art.

Ref./Year	Event	Platform	Keywords	Detection Model
[17] 2020	Four twitter datasets are used. 4709 events posted till January–April 2018 are collected from Snopes.	Twitter, Snopes	-	DL Model
[18] 2020	Six events (German wings, Sydney siege, Ferguson unrest, Ottawa shooting, Boston bombings, Charli Hebdo) has been considered for experimental analysis	Twitter	-	DL Model
[19] 2020	PHEME 5 events data set. Twitter15 and Twitter16 dataset.	Twitter	-	DL Model
[20] 2019		Twitter and Sina Weibo	-	DL Model
[21] 2019	Benchmark Pheme rumor dataset	Twitter	-	DL Model
[22] 2019	A total of 271,000 tweets were collected, consists of 89 events of rumor, and 88 events of non-rumor	Twitter	-	TML Model

[23] 2019	Zubiaga[24] and Kwon[25] dataset	Twitter	-	TML Model
[26] 2019	Weibo and Twitter. There are 2,313 and 2,351 events belonging to rumor and non-rumor in Weibo, 498 and 494 events belonging to rumor and non-rumor in Twitter, respectively	Twitter	-	DL Model
[27] 2019	Six events (German wings, Sydney siege, Ferguson unrest, Ottawa shooting, Boston bombings, Charli Hebdo) has been considered for experimental analysis.	Twitter	-	Hybrid Learning Model
[28] 2019	Four publicly available data sets covering a wide range of real-world events. SemEval-2015 task 1 data, PHEME data set, CrisisLexT26, Twitter event data, CREDBANK	Twitter	-	DL Model
[29] 2019	4664 events from Sina Weibo dataset and 992 events from the Twitter dataset provided by [30]	Twitter	-	DL Model
[31] 2019	5 different controversial topics[14].	Twitter	Obama, air France, cell phone, UK Riots.	TML Model
[32] 2019	Six events (German wings, Sydney siege, Ferguson unrest, Ottawa shooting, Boston bombings, Charli Hebdo) has been considered for experimental analysis	Twitter	-	TML Model
[33] 2019	Reedit rumor dataset	Snopes and Politifact	-	TML
[34] 2018	-	Twitter	-	DL Model
[35] 2018	-	Weibo and Twitter	-	DL Model
[36] 2018	SemEval 2017 rumor detection dataset has been employed.	Twitter	-	DL Model.
[37] 2018	Two standard real-world datasets Twitter 15 and Twitter 16	Twitter	-	DL Model
[38] 2017	-	Sina-weibo	-	DL Model
[39] 2017	-	Snopes	-	DL Model
[40] 2017	-	Snopes	-	Hybrid Model.
[41] 2016	Sports, political, local artistic news. etc.	Collaborative Framework(Google, ANN, BBC,	-	TML Model

		Wattan and Al-Jazeera)		
[42] 2015	Terrorism such as the Boston Marathon Bombing and Garden hose.	Twitter	-	TML Model
[43] 2014	Six well-known events of 2013(Boston Marathon Blasts, Typhoon Haiyan, cyclone phailin, Washington navy yard shooting, polar vortex cold wave, Oklahoma tornadoes).	Twitter	-	TML Model
[44] 2014	Twenty news event-related topics occurred between 1 June 2013 and 15 October 2013 based on current news.	Twitter	US government shutdown, Iran-US relationship, Sarin attack in Syria confirmed, Shipwrecked at Europe, Egypt state of emergency, Train kills dozens in India, etc.	TML Model
[45] 2014	Thematic data(related to specific topic).	Sina Weibo	MH370, Malaysia Airlines, Losing Contact, Malaysia, Black Box, Pray, Crash, Alien, Nanning). The experiment has been done on the workstation with a 12-core CPU of Intel Xeon E5-2620.	TML Model
[46] 2012	Terrorism, natural disaster.	Twitter	ex-terrorist, terrorism, earthquake, tsunami, UK riots, terrorism event.	TML Model
[47] 2012	Controversial topics	Sina Weibo	Keyword published by Sina Weibo rumor debunking service.	TML Model
[48] 2011	Trending topic	Twitter	recycle, earth, save, reduce, reuse, etc.	TML Model
[14] 2011	5 different controversial topics.	Twitter	Obama, air France, cell phone, UK Riots.	TML Model

Table 2.2. List of the dataset that has been adopted by state-of-the-art for detecting misleading content.

Dataset	Key Features	Objective	Type of Data
PHEME [24]	The dataset contains the (1,972) rumors and non-rumors (3,830) including five news breaking stories.	Rumor Detection	Text
KWON [49]	The dataset contains 47 events of rumors and 55 events of non-rumors from Twitter.	Veracity Assessment.	Text

Medieval [50]	The Dataset contains 9000 rumors tweets and 6000 non-rumors tweets based on 17 events.	Detect False Multimedia Content.	Text and Images
RUMDECT [30]	The dataset contains two types of data from Weibo and Twitter. The Weibo data contains 2313 rumors and 2351 non-rumors collected from Sina Weibo rumor debunking service and for Twitter data events that occurred during March-December 2015 are collected from rumor debunking service Snopes. Overall 778 events have been reported during this period.	Rumor Detection	Text
RUMOUREVAL [51]	The dataset is created for Rumor Evaluation 2017, where 297 rumors threads are collected including 297 source and 4,222 reply tweets.	Veracity Assessment.	Text
MULTI [52]	The dataset released in 2017 having 4749 posts of rumor and 4779 posts of non-rumor are collected from Weibo official rumor debunking service. First multimodal dataset released to detect rumors on the Weibo platform which include textual as well as visual data.	Rumor Detection	Text
CrisisLexT26 [53]	The dataset incorporates tweets related to 26 hazardous events take place from 2012 and 2013	Rumor Detection	Text
SNAP data [54]	The dataset comprises of 476 million tweets collected between June and December 2009.	Rumor Detection	Text
[55]	Dataset is the collection of multimedia data having 50,287 tweets and 25,953 images in fake and real news events. • 23,456 fake news tweet • 26,257 real news tweet • 10, 231 fake news images • 15,287 real news images	Detect False Multimedia Content	Text and Images
[56]	Dataset is the collection of medium sized images (1000x700 or 700x1000). The dataset is subdivided into multiple datasets (D0, D1, and D2) • D0 dataset is composed of 50 not compressed images with simply translated copies • Dataset D1 is created by copy-pasting objects after rotation. • Dataset D2 is created by applying scaling to the copies.	Detect False Multimedia Content	Images
Boididou, C *et al.* [33]	Data is collected from the list of 17 news stories events in VC MediaEval 2015 containing 193 real images, 218 fake images, and two cases of misused videos.	Hoax Detection	Images and Videos
Vosoughi, S *et al.* [34]	Dataset is comprised of 209 rumors, including 938,806 tweets, concerning real-time events.	Rumor Detection	Text

Shu *et al.* [36]	FakenewsNet dataset contains 211 fake news and 211 true news that is gathered from BuzzFeed.com and PolitiFact.com	Fake News Detection	Text
Horne *et al.* [37]	The BuzzFeed election dataset contain 36 real news stories and 35 fake news stories during the 9-months before the US Presidential Election.	Fake News Detection	Text
[57]	Dataset is the collection of 12,836 short statements along with its veracity. The dataset is labeled with a discrete set of values from 1 to 6 corresponding to pants-fire, false, mostly-false, half-true, mostly-true, and true.	Fake News Detection	Text
[58]	Fake Video Corpus(FVC) is the dataset[12] is the collection of 381 videos in which 201 are fake and 180 are Non-Clickbait and the FVC 2018 is the collection of 1,675(Fake) and 993(Real), samples[13].	Clickbait video detection	Video
Constraint-2021 COVID-19 Fake News Detection Dataset[14]	The dataset is the collection of 5600 real samples and 5100 misleading samples.	Fake news detection	Text
Clickbait challenge dataset[15]	The dataset is the collection of 5523 clickbait and 16474 Non-Clickbait.	Clickbait detection	Text

2.2 Feature used for FC detection

Feature extraction is one of the crucial phases of any machine learning model, which plays an important role in accurate classification. Some of the commonly adopted features for FC detection by existing state-of-the-art methods on multimedia data with its detailed descriptions are shown in Table 2.3 and 2.4 respectively. In Table 2.3, we have provided some recent works, where 15 different prominently used features categories (Message-based, User-based, Topic-based, Propagation-based, Content-based, Network-based, Twitter-based, Linguistic, Temporal, User-behavioural, Diffusion, Structural, Social, Visual and Statistical Features) corresponding to feature no $F1, F2\ , F3, F4 \dots\dots\dots F15$ is considered for FC analysis on multimedia data. The detailed description of each set of features belonging to a different category is discussed below in Table 2.4. It can be noted from Table 2.3, that most of the works have employed content $(F2)$ and user-based $(F5)$ features compared to other features for false

[12] https://github.com/MKLab-ITI/fake-video-corpus/blob/master/FVC.csv
[13] https://github.com/MKLab-ITI/fake-video-corpus/blob/master/FVC_dup.csv
[14] https://constraint-shared-task-2021.github.io/
[15] Clickbait Challenge (webis.de)

information detection, and their usage is still in a continuation. This reflects that they are found to be effective for the FC detection task. It can also be proved from the research findings shown in Table 2.3. Additionally, it is observed from [59],[60],[61],[62],[24],[14], that content-based features play a major role in false information detection. The authors of [14], reports that hashtags used in rumorous tweets differ from other tweets. On the other hand, hashtags used by people who believe and disseminate rumors are different from the people who deny it. It has been reported that the URL plays a crucial role in information diffusion. Features that are calculated using the content language model are efficient in attaining high precision and recall. Whereas, user-based features are also prominently employed for this task. The authors of [60], reports that retweet ratio is an efficient clue which can indicate that a rumor is spreading. The normal retweet ratio is 8.03 % if it is more than that it's an indicator of rumor.

Table 2.3. Handcrafted features mostly used by earlier state-of-the-art

Reference	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
[63] 2020														✓	✓
[64] 2019		✓		✓	✓					✓					
[65] 2019					✓								✓		
[2] 2017		✓						✓	✓	✓			✓		
[55] 2017														✓	✓
[25] 2017		✓						✓	✓	✓		✓			
[59] 2015					✓	✓	✓								
[49] 2013								✓	✓			✓			
[46] 2012		✓			✓						✓				
[14] 2011					✓	✓	✓								
[48] 2011	✓	✓	✓	✓											

Table 2.4. Handcrafted Features used for Rumor Analysis on Multimedia data.

Feature No	Feature Category	Feature Name
F1	Message-based	Length character
F1	Message-based	Length words
F1	Message-based	Contain question mark
F1	Message-based	Contain exclamation mark
F1	Message-based	Number of URLs.
F1	Message-based	Contain user-mention

F1	Message-based	Contains hashtag
F1	Message-based	Is Retweet
F1	Message-based	Sentiment positive words
F1	Message-based	Sentiment Negative Words
F2	User-based	Registration age
F2	User-based	Statuses count
F2	User-based	Count Followers
F2	User-based	Count Friends
F2	User-based	Is Verified
F2	User-based	Has Verified
F2	User-based	Has Description
F2	User-based	Has URLs
F2	User-based	Count Followers
F2	User-based	Time interval of the last two tweets.
F2	User-based	Total count of tweets per day.
F2	User-based	Author registration age.
F2	User-based	Count of Followers
F2	User-based	Count of Followees
F2	User-based	Is a verified user
F2	User-based	The count of tweets the user has authored in the past.
F3	Topic-based	Tweet count
F3	Topic-based	The average length of the tweet
F3	Topic-based	The fraction of tweet having a question mark
F3	Topic-based	The fraction of tweets having URLs.
F3	Topic-based	Count of the distinct author of tweets.
F3	Topic-based	Count of the distinct author of tweets
F3	Topic-based	Count of positive words
F3	Topic-based	Count of negative words
F4	Propagation-based	The degree of the tweet in the propagation tree
F4	Propagation-based	Count of tweets in the largest subtree of the root
F4	Propagation-based	The depth of the propagation tree
F4	Propagation-based	Comment count.
F4	Propagation-based	Count of the post.
F4	Propagation-based	Is repost?.
F5	Content-based	Count of ”@”
F5	Content-based	Count of ”#”.
F5	Content-based	Count of sentiment word
F5	Content-based	Count of URLs
F5	Content-based	Number of words contained in a post
F5	Content-based	Type of client posting a post
F5	Content-based	Representing time interval between user registration and time of posting
F6	Network-based	Feature related to retweet information
F7	Twitter-based	Twitter-based features include features related to twitter such as hashtags, and they likely to investigate whether the hashtags used in rumor are different from non-rumor tweets
F8	Linguistic based	The ratio of Tweets Containing Negation
F8	Linguistic based	Average Formality and Sophistication of Tweets.
F8	Linguistic based	The ratio of Tweets Containing Opinion and Insight
F9	Temporal	A set of distinct features has been observed from the time series data and founds that rumors tend to have multiple and periodic spikes, while non-rumors typically have a single prominent spike.

F10	User-behavioral	Verified user or not.
F10	User-behavioral	Count of followers
F10	User-behavioral	The average number of followers per day
F10	User-behavioral	Count of posts per day.
F10	User-behavioral	Count of possible microblogs sources
F10	User-behavioral	Count of reposts and comments.
F10	User-behavioral	The ratio of Questioned comments
F10	User-behavioral	Count of corrections.
F11	Diffusion based	Time the tweet has been cited
F11	Diffusion based	Time of the original tweets been cited if it is a re-tweet.
F12	Structural based	The features including structural characteristics of the rumor diffusion network.
F13	Social Feature	Count of tweets written by the author
F13	Social Feature	Count of lists that include the author's account
F13	Social Feature	The following ratio of the author’s account
F13	Social Feature	The age of the author's account
F13	Social Feature	The account status of the author, whether the account is verified or not.
F14	Visual Feature	Visual Clarity Score
F14	Visual Feature	Visual Coherence Score
F14	Visual Feature	Visual Similarity Distribution Histogram
F14	Visual Feature	Visual Diversity Score
F14	Visual Feature	Visual Clustering Score
F15	Statistical feature	Count: Set of all images presents in a news event.
F15	Statistical feature	Image Ratio: Ratio of the image tweets in the set of all tweets.
F15	Statistical feature	Image Ratio II: It is the ratio of image number to tweet number
F15	Statistical feature	Multi-Image Ratio: It is the ratio of multi-image tweets in the set of all tweets.
F15	Statistical feature	Multi-Image Ratio II: It is the ratio of the multi-image tweet in the set of all image tweets.
F15	Statistical feature	Hot Image Ratio: The ratio of the most popular image in the set of distinct images.
F15	Statistical feature	Long Image Ratio: It is the ratio of tweets with long images in a set of all image tweets.

2.3 Detection of Fraudulent content in OSNs

This section covers a detailed description of false information detection methodologies employed under TML and DL methods based on multimedia data. Each model has been discussed in detail, including various measures (performance, classification model, the methodology employed, aim, and future directions). The FC detection problem can be defined as, the set of social media posts is given as an input $p = p1, p2 \dots \dots .p(n)$, the model needs to identify is each of the given posts pi, is fake or real, usually formulated as a binary classification problem. A detailed description of the methodologies concerning to diverse set of deceptive activities such as Fake News, Rumor, Hoax, and Clickbait on multimedia data as shown in Table 2.5.

Table 2.5. Methodologies adopted by the state-of-the-art for misleading multimedia content

Reference	Year	Method	Performance/model	Aim	Model
[66]	2021	The method based on a semi-supervised classification approach utilize attentions sampled from a Gumbel softmax distribution to distill contexts.	Achieved 97% accuracy	CD	DL
[67]	2021	The features are extracted from text using the ontology method. Content and sentiment features have been used.	Achieved an F-measure of 92%.	CD	TML
[68]	2021	Proposed a novel method of generating stylish headlines from the original data using style transfer to detect clickbait.	Reported 93% accuracy using logistic regression	CD	TML
[69]	2021	Keyword Based Method Distance-based method Neural Network and advanced text processing	Logistic Regression	HD	TML
[70]	2021	Text matching method using Levenshtein distance measure. Bag of words (URGENT, ATTENTION, PLEASE.) Uppercase percentage, keyword percentage.	Rule-based system	HD	TML
[17]	2020	Dual convolutional neural networks are used for processing temporal, structural, and linguistic features of a post.	DCNN outperforms the most recent approaches by 5–35% during the early stages.	RD	DL
[18]	2020	Time series data from the parsed tweets for the different time intervals is given as an input to the ensemble model. GRU,LSTM,RNN	There is an improvement in the classification performance by 7.9% in terms of micro F1 score compared to the baselines.	RD	DL
[23]	2019	Proposed a novel classification approach called OCC in which s the classifier is trained by one class only.	It has been observed that the OCC approach can recognize rumors with a high level of F1 score. Achieve the F1-score of 74% on Zubigaset and 93% on Kwonset.	RD	TML
[64]	2019	SVM, Decision Tree, Random Forest, KNN, GBDT, Xgboost.	Precision:0.827 Recall :0.837 F1-score:0.825 AUC at 0.895.	RD	TML
[71]	2019	Convolution unit in Tree LSTM	Exceeds the best prior work by 12% in f1-score.	VA	TML
[2]	2017	DTW(Dynamic Time Wrapping) and HMM(Hidden Markov Model)	Accuracy(HMM): 0.75	VA	TML

[55]	2017	SVM, Logistic Regression, KStar, and Random Forest.	Accuracy of 83.6%.	RD	TML
[72]	2016	Decision Tree(DT) , Logistic Regression(LR)	Accuracy(DT):96.6% Accuracy(LR): 82.8%	VA	TML
[24]	2016	SVM, CRF, Naive Bayes, Random Forest	Precision(CRF): 0.667	RD	TML
[73]	2016	DT(Dynamic Tree Time Structure, DSTS(Dynamic Series Time Structure), Hybrid DSTS with SVM classifier, Hybrid DSTS with RBF kernel	Accuracy(DT): 0.985	RD	TML
[74]	2015	Decision Tree, SVM, RF	Accuracy(RF): 0.867	RD	TML
[3]	2015	SVM with hybrid kernel including Random walk kernel and an RBF kernel.	Accuracy: 91.3%	RD	TML
[49]	2013	Random Forest, SVM	High precision and recall in the range of 87% to 92%,	VA	TML
[62]	2013	SVM	-	RD	TML
[48]	2011	Decision Tree, Bayesian Network, SVM	Precision and recall in the range of 70% to 80%.	VA	TML
[47]	2012	SVM with RBF kernel	The model has been evaluated on the following measures like precision, recall, f-measure. The classification accuracy has been improved via incorporating two newly proposed features to varying degrees which are 5.42%, 4.73%, and 6.317% into the task of classification.	RD	TML
[75]	2015	J48 classifier	F-measure of more than 0.82	RD	TML
[40]	2017	CNN,RNN,LSTM		RD	Hybrid
[76]	2013	Decision Tree, Naïve Bayes	Decision Tree: 97%	Fake Images detection	TML
[55]	2017	To characterize fake and real news event image distribution patterns, several visual and statistical features have been proposed. Firstly introduced image-based features for fake news detection	SVM, Logistic Regression, KStar, and Random Forest. Accuracy of 83.6%.	Manipulated images detection	TML
[50]	2014	To classify whether the post associated with an image is fake or real. The features related to post, user account and the image has been extracted.		Manipulated images detection	TML

[56]	2015	A novel hybrid approach (halfway between block and point-based methods) has been proposed for copy-move forgery detection that relies on the analysis of local key points of triangles. Various inner features like color, geometrical properties(angles), and vertices that compose the triangles have been analyzed. The point of interest has been extracted from an image using SIFT, SURF, and Harris .		Manipulated images detection (Copy move forgery).	TML
[77]	2019	Proposed an algorithm that verifies the veracity of the image text via inspecting it on the web and then verifying the authenticity of the top 15 google search results. Reality parameter (RP) has been proposed for classifying an event as real or fake.		Fake news detection	TML
[37]	2018	Modeled the structure as a propagation tree to categorize and segregate rumorous and non-rumorous claims based on a comparison study of their tree-based similarities. For rumor representation and learning two recursive neural models has been proposed based on top-down and bottom-up tree-structured neural network.		Early Rumor Detection	DL
[39]	2017	A novel provenance-aware approach based on RNN has been proposed. – Provenance and text information of the post are combined to enhance the accuracy of the rumor prediction system. To capture the temporal dependency between the posts, the RNN based model has been employed. However, to capture the long-term dependency of input, the author has used LSTM instead of vanilla RNN.	The model outperforms other baselines with an accuracy of 0.85(Improvement by 9%) and a recall of 92% improvement by 22%.	Rumor detection	DL

2.4 Research Gaps

From the earlier study as we discussed in the previous sections concerning to diverse set of deceptive activities, we come up to the following research gap as given below.

- The case when videos presenting false information as true and user responses are in support of the claim has not been addressed.
- None of the previous studies have provided the analysis w.r.t multiple web search engines for the evidence collection.
- Lack of FC detection techniques over Cross/Multi-Web Platforms.
- Lack of publicly available data towards the development of FC detection. Especially in the form of images, audio, and videos.
- Very few studies have been explored the concept of detecting FC incorporating text embedded images.
- Among other techniques, Hoax and Clickbait are the least addressed area, which requires further attention.
- Few studies are reported concerning ensemble-based learning for the detection of fraudulent content.
- Video content-based features are not explored as a clue measure for the verification of misleading videos.

2.5 Research Objectives

From the research gaps that we have discussed in the earlier section, we come up with the following research objectives.

- To develop an effective technique for detecting deceptive claim on web platforms.
- To develop an effective technique for detecting deceptive claim accompanying multimedia content on web platforms.
- To develop a Multi-Web Platform framework for detecting deceptive claim.
- To know the effectiveness of the novel algorithm, a comparative study and implementation is to be conducted.

Each of the given objectives has been detailed discussed in the following section. The first objective presents the proposed technique for detecting deceptive claims on web platforms and its evaluation in detail. The next objective incorporated the concept of detecting fraudulent claims accompanying multimedia content. In this work, we have mainly considered images and videos concerning multimedia content. Whereas, in the third objective, we aim to develop a multi-web platform framework for detecting deceptive claims on the social media platforms and the last objective is covering the comparative study of the proposed methods with the other state-of-the-art method.

Chapter 3

Detection of Fraudulent Claims on Web Platforms

This chapter explained the proposed methodology to detect fraudulent content on social media available in the form of text. A detailed description of the problem statement, data collection, the feature extraction process, and the methodology adopted has been provided in this chapter. The effectiveness of the proposed approach is explained as well as validated through experiments on standard datasets and a state-of-the-art comparison study of the obtained results has been provided.

This chapter covers the proposed techniques for detecting deceptive claims on web platforms. Due to the widespread use of social media platforms and their easy services for information sharing, people freely share their thoughts and opinions among the public regarding any event. The information can be shared in many different forms either they are completely text format, or they are in the form of text embedded images, or text accompanying with some multimedia content (i.e. images or videos). The chapter will describe in detail the technique employed for the prediction of FC having only text as an input. An overview of the complete model has been detailed discussed in the following paragraph. The techniques we covered here in this concern are based on two ways. The input is given as text embedded images or in the simple text format. In the first case, the input is considered as text embedded images where text is embedded in an image, and using the OCR technique (optical character recognition) the text content is retrieved from an image. The second case is when the post itself is a text content. These two cases have been considered in this chapter and techniques involved in each of these cases have been discussed in detail.

3.1 Detecting Fraudulent content: Text Embedded Images

In this case, text-embedded images are given as an input to the system. The OCR technique is utilized to process text embedded images in order to extract text from an image and then the processed text further be utilized to retrieved the evidential clues for the prediction of false information. we propose a model which is concerned with the veracity analysis of information on various social media platforms available in the form of images. It involves an algorithm that validates the veracity of image text by exploring it on the web and then checking the credibility of the top 15 Google search results by subsequently calculating the reality parameter (Rp), which if exceeds a threshold value, an event is classified as real else fake. To test the performance of our proposed approach, we compute the recognition accuracy, and the highest accuracy is compared with similar state-of-the-art models to demonstrate the superior performance of our approach.

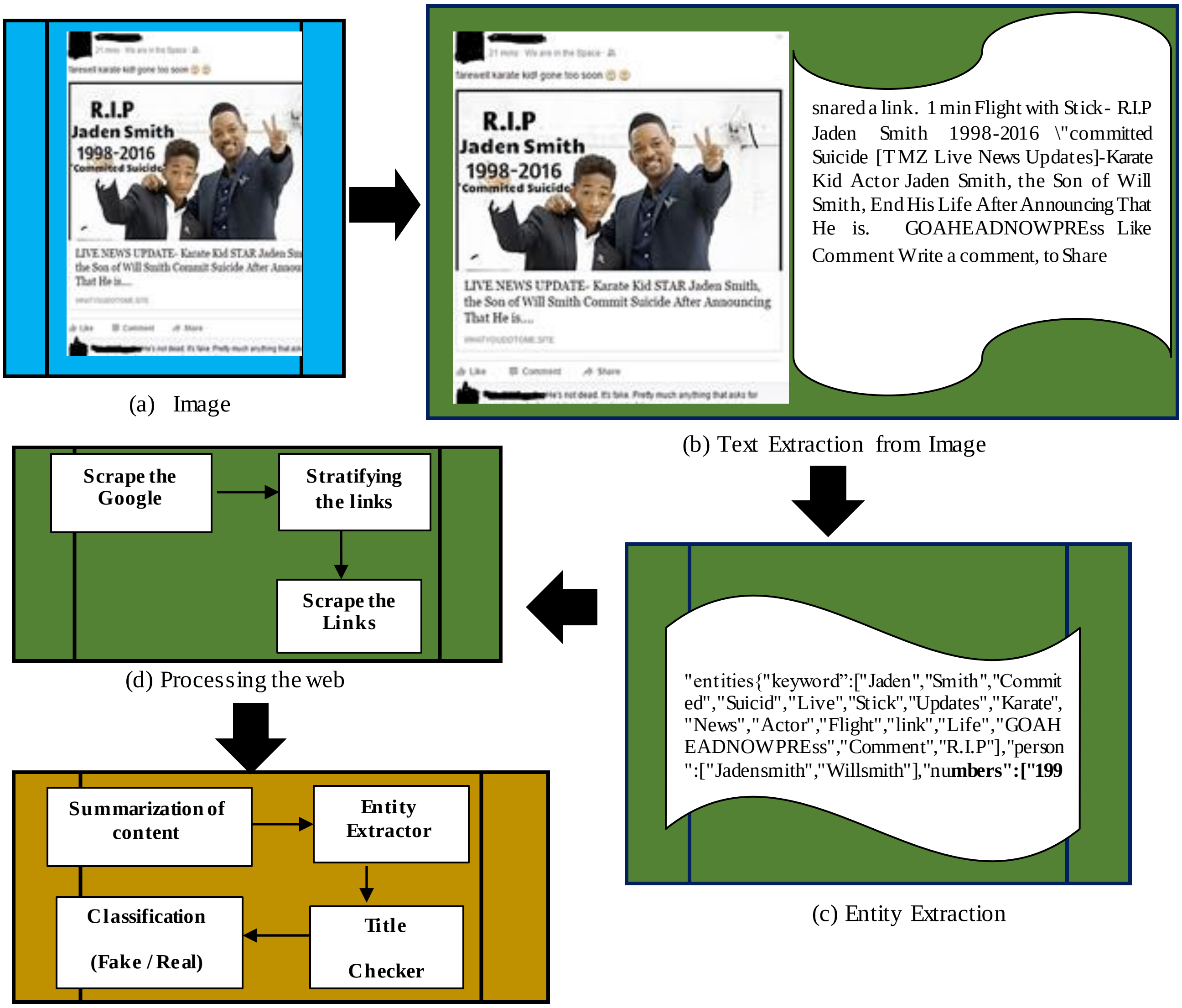
The flow diagram of the proposed algorithms is shown in Fig 3.1.1. It can be seen how all smaller units are connected, and finally, an overall system is developed. The purpose of the proposed system is to analyze the veracity of the news events that are floating in the form of images in social media. The framework is composed of four basic units: (i) Text extraction from image (ii) Entity extractor (iii) Processing the Web (iv) Processing Unit. In the following section, the details about each module are discussed in depth.

3.1.1 Text Extraction from Image

The input image goes through a series of transformations which facilitates the process of text extraction. The first module, "Text Extraction from Image," performs the function of extracting text from the image. Here, we use the method proposed in [78], for detection of text region and then with the help of optical character recognition (OCR), the text is extracted from images. The key steps of the algorithm are: Firstly, Maximally Stable Extremal Region (MSER) detection is used to locate every text location comprised of text with various fonts and sizes. Secondly, Maximally Stable Extremal Region (MSER) enhancement is used to make the boundaries of letters more identifiable. Thirdly, a Stroke Width Detector is applied for the detection of the stroke width of characters. For this purpose, ray vectors are calculated using Eq.3.1.1. Finally, filtering is done to eliminate the area which is unlikely to contain text characters. Text region components recognized in the above process go through the Optical Character Recognition module to recognize text from the image. In this work, we considered only the English language.

$$Rx = Px + ncos(\theta), \quad Ry = Py + n \sin(\theta) \tag{3.1.1}$$

where *ray vector* $\mathrm{r} = [\mathrm{R_x}, \mathrm{R_y}], \mathrm{p} = [\mathrm{P_x}, \mathrm{P_y}]$ represents the boundary pixel position, n is the iteration index, and θ represents the gradient direction. Eq. 1 is used to find the boundary pixel in the edge image by increasing the value of index $n.$ If the direction of two boundary pixels is the opposite, then the ray is included, else discarded.



(a) Image

snared a link. 1 min Flight with Stick- R.I.P Jaden Smith 1998-2016 \"committed Suicide [TMZ Live News Updates]-Karate Kid Actor Jaden Smith, the Son of Will Smith, End His Life After Announcing That He is. GOAHEADNOWPREss Like Comment Write a comment, to Share

(b) Text Extraction from Image

"entities{"keyword”:["Jaden","Smith","Commited","Suicid","Live","Stick","Updates","Karate","News","Actor","Flight","link","Life","GOAHEADNOWPREss","Comment","R.I.P"],"person":["Jadensmith","Willsmith"],"nu**mbers":["199**

(c) Entity Extraction

(d) Processing the web

(e) Processing Unit

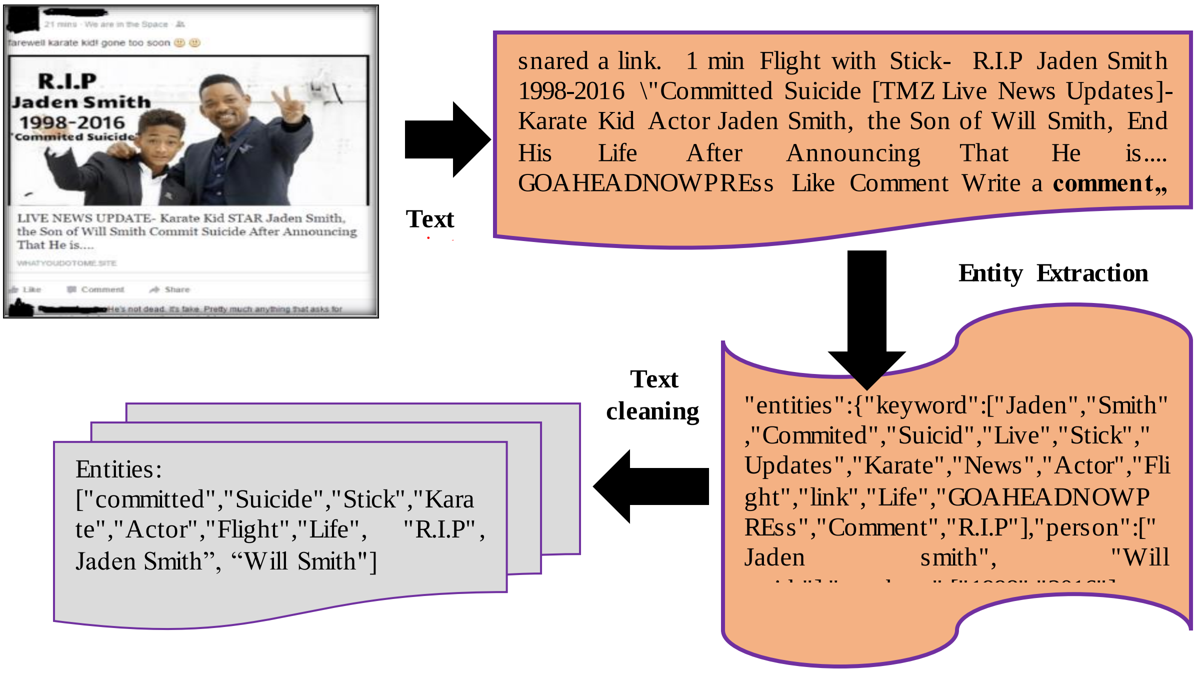
Fig 3.1.1. Underlying Architecture of the proposed system.

3.1.2 Entity Extractor

The Second module, “Entity Extractor,” is responsible for the extraction of entities from text. In the following section, the process of entity extractor is explained in brief. Initially, the extracted text from the image is processed to fetch the various entities from it. In the next step, each entity goes through the process of text cleaning, which is further responsible for the following: 1) striping of all non-alphabetic characters, 2) removing multiple occurrences of the words, 3) checking whether the word is a valid English word, 4) checking each word for spelling errors with 1-edit distance, 5) removal of any media house name or newspaper name to remove bias. Fig 3.1.2 shows a working example of entity extraction.

Algorithm 3.1.1

```
procedure ProcessTheLink (text)
    L← GetGoogleSearchResults (text)
    R← GetReliableLinks (L)
    For each link in R
        Content ←ScrapeLink (link)
    return Content
end procedure
```



[redacted]
21 mins · We are in the Space
farewell karate kid! gone too soon 🙁 🙁
R.I.P
Jaden Smith
1998-2016
"Commited Suicide"
LIVE NEWS UPDATE- Karate Kid STAR Jaden Smith, the Son of Will Smith Commit Suicide After Announcing That He is....
WHATYOUDOTOME.SITE
Like Comment Share
[redacted] He's not dead. It's fake. Pretty much anything that asks for

snared a link. 1 min Flight with Stick- R.I.P Jaden Smith 1998-2016 \"Committed Suicide [TMZ Live News Updates]- Karate Kid Actor Jaden Smith, the Son of Will Smith, End His Life After Announcing That He is.... GOAHEADNOWPREss Like Comment Write a **comment,,**

"entities":{"keyword":["Jaden","Smith","Commited","Suicid","Live","Stick"," Updates","Karate","News","Actor","Flight","link","Life","GOAHEADNOWPREss","Comment","R.I.P"],"person":[" Jaden smith", "Will

Entities: ["committed","Suicide","Stick","Karate","Actor","Flight","Life", "R.I.P", Jaden Smith", "Will Smith"]

Fig. 3.1.2. Entity extractor Process

3.1.3 Processing the Web

In the third module, the string of extracted entities is searched on Google, and links to the search result are collected, which are further scraped for their content. Hence, in this module, the Google search results are fetched and are further categorized into a reliable or unreliable link, followed by scraping the content of the reliable links. To this aim, web scraping techniques are applied. Algorithm 3.1.1 describes the process, and it returns the dictionary with links as the key, and content as the value. The following steps performed in this module are described in detail:

- Scraping Google is the first step in this module, which searches a specific string on the Google search engine and scrapes the results. To this aim, we use selenium, a portable software testing framework for web applications.

- Stratifying the links is responsible for classifying the links into reliable or unreliable based on a list compiled by us.
- Finally, scraping the links is concerned with scraping the content of each reliable link. To this aim, we use the already-available tool scrappy, a framework written in Python for the same.

3.1.4 Processing Unit

The fourth module, "Processing unit," is the final module in the proposed system and is concerned with the classification of the event. The classification is done into two categories "fake" or "real". In the following text, sub-modules are described with their working:

- Summarization of content is concerned with producing a concise summary of the content of the web pages. To help us in this process, we use Python's natural processing tool (NLTK). For obtaining a summary, the extraction parameter has been tuned to suit our needs. Based on the values obtained at each iteration, the best results were achieved by limiting the content of the summary to 40-65 words. Moreover, only reliable links go through this process.
- Entity extractor is concerned with the extraction of entities from the summary produced by summarization of content.

Title Checker is responsible for calculating the percentage match of entities extracted from the image with that of the title of each reliable link. Furthermore, if a match is above a threshold value and a particular keyword is found in the title, then the link is not considered reliable.

Algorithm 3.1.2 describes the title checker process where three parameters (title, query, bag of words) are passed to the title checker procedure.

Algorithm 3.1.2: Title Checker

```
procedure TitleChecker (title, query, bag of words)
    P←percentage match (title, query)
    If P > 0.30 and the title has "bag of words"
        return False
    Else
        return true
    End If
End procedure
```

Classification is the submodule where the final decision is taken based on the value of the reality parameter. The content of the reliable links goes into a summarizer, which summarizes the content. The summary goes through entity extractor, and then the percentage match is found between extracted entities from summary with query searched on Google. If the match comes out to be zero then the link shift to an unreliable link, else it stays as reliable. Furthermore, the content of the reliable link further goes through the title checker, which if return false, the link is considered as unreliable. The number of reliable and unreliable links is then used to find the reality parameter, which is used to classify news as fake or real. Reality parameter value has been calculated using Eq. 3 to classify the news as fake or real. If Rp >=40 then the news is classified as real, else it is classified as fake. Algorithm 3.1.3 describes the final module:

Algorithm 3.1.3: Fake News Classification

```
procedure Classifying Real And FakeNews (reliableLinks, Content, query, totalLinks)
        totalReliableLinks = 0
For each link in reliableLinks
        If (TitleChecker (Content [link]) and SummaryMatch (Content [link], query)> 0.0)
                        totalReliableLinks = totalReliableLinks + 1
                End If
        End For
        Rp = (totalReliableLinks / totalLinks) * (100)
        If Rp >= 40
                Classify as real
        Else
                Classify as fake
        End If
End procedure
```

The credibility of multimedia content on social media is a new and emerging problem, and there are very limited datasets available for the classification of an image as fake or real. One of the publicly available datasets (PHEME) of rumors and non-rumors is provided by [24]. The dataset includes a collection of 1,972 rumors and 3,830 non-rumors associated with five breaking news stories from Twitter. The authors of [79], collected a dataset from the VC-MediaEval 2015 task, which consists of tweets related to 17 hoaxes including 193 real images, 218 fake images, and two cases of misused videos. Another dataset related to rumor and non-rumor is provided by [2]. This dataset contains 209 rumors, including 938,806 tweets concerning real-time events (2013 Boston Marathon bombings, 2014 Ferguson unrest, the 2014 Ebola epidemic, etc.). Table 3.1.1 describes some of the publicly available datasets for fake news detection. Moreover, Zhang et *al.* [80], presented a survey that discussed the publicly available datasets for fake-news analysis.

Table 3.1.1. Dataset details for fake news analysis.

References	Dataset Description	Input Data Type
Zubiaga *et al.* [24]	PHEME dataset includes a collection of 1,972 rumors and 3,830 non-rumors associated with five breaking news stories.	Text
Boididou, C *et al.* [79]	Data is collected from the list of 17 news stories events in VC MediaEval 2015 containing 193 real images, 218 fake images, and two cases of misused videos.	Text and Images
Vosoughi, S *et al.* [2]	Dataset is comprised of 209 rumors, including 938,806 tweets, concerning real-time events.	Text
Shu *et al.* [81]	Fake news Net dataset contains 211 fake news and 211 true news that is gathered from BuzzFeed.com and PolitiFact.com	Text
Horne *et al.* [82]	The BuzzFeed election dataset contains 36 real news stories and 35 fake news stories during the 9-months before the US Presidential Election.	Text

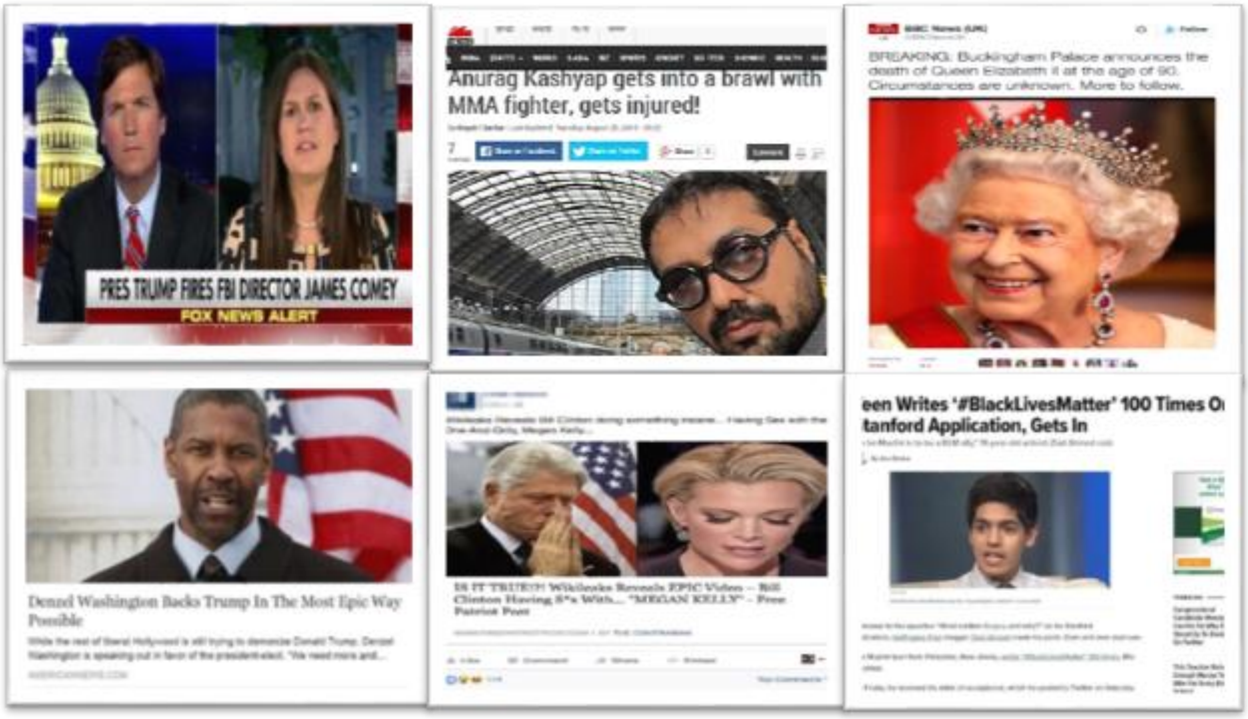

PRES TRUMP FIRES FBI DIRECTOR JAMES COMEY
FOX NEWS ALERT
Anurag Kashyap gets into a brawl with MMA fighter, gets injured!
BREAKING: Buckingham Palace announces the death of Queen Elizabeth II at the age of 90. Circumstances are unknown. More to follow.
Denzel Washington Backs Trump In The Most Epic Way Possible
een Writes '#BlackLivesMatter' 100 Times O
tanford Application, Gets In

Fig 3.1.3. Sample images of the dataset

Hence, to test the performance of the developed algorithm, a dataset of thousands of images are collected from Google images, the Onion, and Kaggle maintaining a balance between fake as well as real news images. The sample images of datasets are as shown in Fig 3.1.3. The news is further divided into three categories: 1) Local/regional news, 2) National news, 3) International news. The cases per category are shown in Table 3.1.2. For the list of verified content producers, we have manually compiled the maximum possible name into a common file.

Table 3.1.2. Category-wise distribution of images

S.NO	News Category	Cases per Category (Total_cases=1000)	#Fake images	#Real images
1	Local/Regional News	150	70	80
2	National News	350	170	180
3	International News	500	250	250

To calculate the optimum number of links required to classify the image, we conducted a series of iterations. Table 3.1.3 shows the result of those iterations on all three types of news, i.e., regional, national, and international. From Table 3.1.3, we choose an optimal value for the number of links to be 15.

Table 3.1.3. Variation in the number of reliable links

NLC	NRLFNE	NRLRNE
0-5	1-3	2-5
6-10	-	2-10
11-14	1-5	2-11
15	1-5	2-12
16	1-6	2-12
17	1-6	2-12
18	1-6	2-13
19	1-6	2-13
20	1-6	2-13

*NLC: Number of Link under consideration, NRLFNE: Number of reliable links for fake news events (min-max), NRLRNE: Number of reliable links for real news events (min-max)

To calculate the effective value of reality parameter (Rp), we conducted a set of iterations on our dataset for different values of reality parameter. Table 3.1.4 shows how the system performs on different values of Rp for national and international news. In the case of local news, the system is not able to classify local news events as real or fake, as most of the prominent content producers do not cover this news. To calculate the accuracy of the system, we have used Eq. 3.1.2.

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} * (100) \qquad (3.1.2)$$

Where TN, TP, FN, and FP are the True Negative, True Positive, False Negative, and False Positive, respectively. In our analysis, we also found that for some cases, the value of reality parameter comes out to be exactly 40% as the Google search index of unreliable sites is better than the reliable sites, and they have a better rank on Google search result. This resulted in the inclusion of those unreliable sites in our set of links, thus shifting out the reliable links from our set. A detailed analysis of factors affecting Google search results is listed in [83]. To calculate the reality parameter, we have used Eq.3.1.3. The proposed system gives the best accuracy at 40%.

$$Rp = \frac{Number\ of\ reliable\ links}{Total\ number\ of\ links} * (100) \qquad (3.1.3)$$

Table 3.1.4. Effect of Reality Parameter on Accuracy

Rp	>= 80	>= 66	>= 53	= 46	>= 40
ANIN	68	73	77	78	85

*Rp: Reality parameter (Percentage), ANIN: Accuracy on National and International News (Percentage).

During our experimentation, we have found that sometimes, credible media sources cover the fake news stating it as a hoax to aware the readers. Table 3.1.5 shows how different fake news is covered by credible news media, but their context is different, and their purpose is to make their readers aware of the fake news events. We have analyzed the news titles and found that there exists a pattern of specific keywords in the title encountering fake news. One of the fact-checking websites Snopes.com[16] has been analyzed to identify the pattern of news titles covering the fake news. We have manually extracted the list of 10 frequently used keywords and have included them in our bag of words. The keywords are Hoax, Hoax fools, False News, Fake, Fake news, Fake death, Rumors, False, Death Hoaxes, Falsely. These bags of words have been incorporated into our algorithm to remove any false positives we might get due to such cases.

The Table 3.1.6 shows the percentage match of page title with news type (national and international news). The minimum value of the title match is chosen as the threshold which is set to 0.30. News events include both fake and real events. Cases, where a 0% match is found, are not shown in the table for both fake (cases where news is covered by credible sources to alert their readers) and real news.

Table 3.1.5. Fake News covered by credible media to aware users

List of news	Title found	Keyword
Brad Pitt died	Brad Pitt is NOT dead as vile online hoax fools fans with virus amid divorce from Angelina Jolie	Hoax fools

[16] https://www.snopes.com/fact-check/celebrity-death-hoaxes/

Women killed, Black lives matters	The story of 19 white women killed by Black Lives Matter supporters is fake news	Fake news
Women defecating boss	How fake news story of woman defecating on boss' desk after hitting $3m jackpot fooled thousands after going viral	Fake news
Facebook to start charging	Will WhatsApp and Facebook start charging? The latest scam tells users to pass on chain messages to avoid costs	Latest scam, scam
Jaden Smith died	Jaden Smith is still not dead after vile 'suicide' hoax continues to baffle fans online	Hoax

TABLE 3.1.6 Title match with news

Type of News	National news (Min-Max)	International news (Min-Max)
Title Match (%age)	0.31-0.71	0.30-0.64

Sometimes due to miscommunication, or under stress or excitement, credible media or newspapers also publish fake news. But after a while, realizing their mistake, they take down the respective web page. But by the time they do that, Google has already crawled and indexed their page, and it starts showing in the Google search results. Now when we search any of such events, it gives us a credible source as a result, but there is no relative content present in that link. Due to this reason, we kept a check on summary matches to be more than zero. Table 3.1.7 shows the summary match for real news.

Table 3.1.7 Showing Summary Match Results

Type of News	Local news	National news (Min-Max)	International news (Min-Max)
Summary Match (%age)	-	0.274 - 0.485	0.23 – 0.41

3.1.5 Effectiveness of fake news detection

To evaluate the effectiveness of our proposed method, a comparative analysis is outlined with three state-of-the-art methods. The performance of the proposed algorithm is measured on the datasets used by different state-of-the-art [24][81][82] as shown in Table 3.1.8. The evaluation metrics used for measurements are Accuracy (Acc), Precision (P), Recall (R), and F1 scores. All the experiments were performed using five-fold cross-validation settings, where the final accuracy is computed by averaging the values across each of the five folds.

Table 3.1.8 Comparison of the proposed approach on state-of-the-art datasets

Method	Name of dataset	Input Type	Classifier	Acc (%)	F1 (%)	P (%)	R (%)
[24] (2016)	PHEME	Text	CRF	-	60.7	66.7	55.6
[81] (2017)	FakenewsNet: 1) BuzzFeed PolitiFact	Text	SVM	86.4 87.8	87.0 88.0	84.9 86.7	89.3 89.3
[82] (2017)	BuzzFeed election	Text	SVM	77	-	-	-
	PHEME	Text	Rule-based	**-**	**69.3**	**73.2**	**65.8**

Proposed Method	1) BuzzFeed 2) PolitiFact	Text	Rule-based	**85.3** **88.0**	**86.77** **88.34**	**85.2** **87.9**	**88.4** **88.8**
	BuzzFeed election	Text	Rule-based	**86**	**-**	**-**	**-**

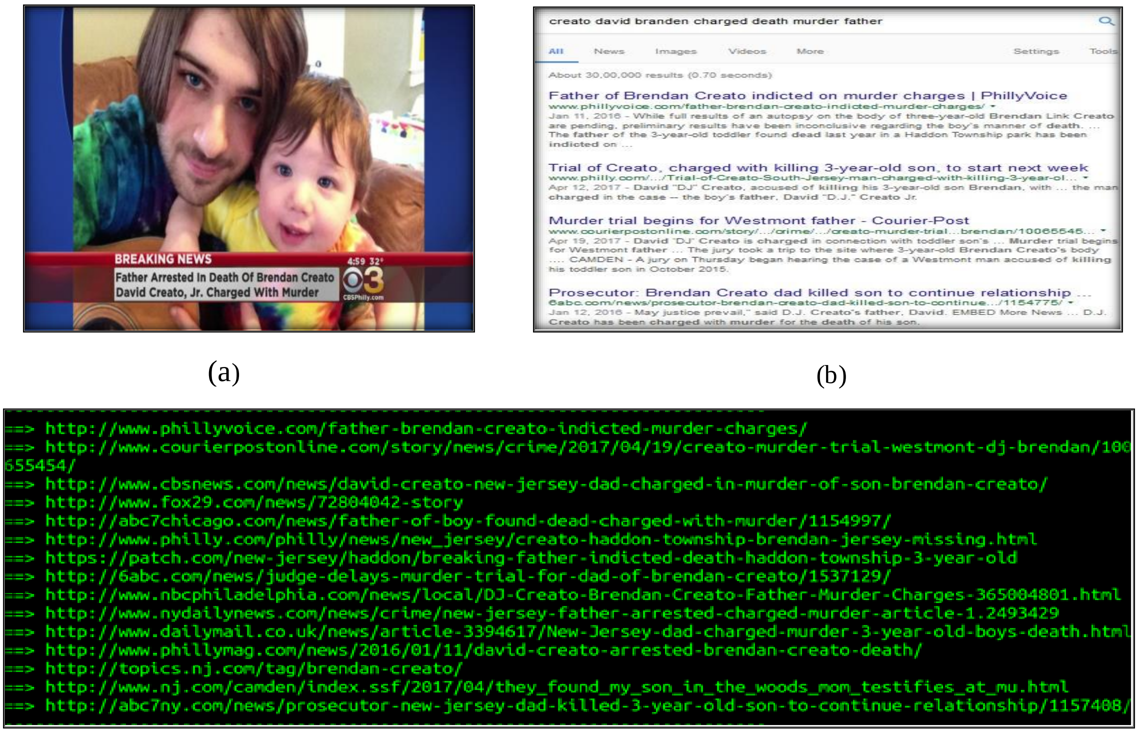
From Table 3.1.8, the evaluation metrics parameter shown in the bold letter are achieved by the proposed algorithms and the value of these parameters are higher than the other state-of-the-arts. Hence, it can conclude that the proposed algorithms outperform in comparison with other state-of-the-arts on similar datasets. The output results are analyzed in terms of the reality parameter (Rp), which is set at 40%, and the number of links is set to 15, giving the highest accuracy. It was observed that the match between the summary of the content of the web page with the search query was ranging from 0 to 70%, whereas the summary was ranging from 40-65 words each. The proposed system outperformed the state-of-the-art system giving an accuracy of 85%. The experimental results show that the proposed system does not perform well on a local/regional news event.

3.1.6 Results

To test the performance of the proposed algorithm, an online test has been conducted on fake news. The sample output on real news and fake news is as shown in Fig. 3.1.4 and Fig. 3.1.5. Table 3.1.9 shows how reality parameter reacts to local, national, and international news. The output results are analyzed in terms of the reality parameter (Rp), which is set at 40%, and the number of links is set to 15, giving the highest accuracy. It was observed that the match between the summary of the content of the web page with the search query was ranging from 0 to 70%, whereas the summary was ranging from 40-65 words each. The proposed system outperformed the state-of-the-art system giving an accuracy of 85%. The experimental results show that the proposed system does not perform well on a local/regional news event.

TABLE 3.1.9. Reality parameter vs. news type

Type of News	Local news	National news	International news
Reality Parameter (Rp)	0.0-0.20	0.26-0.80	0.46-0.86


BREAKING NEWS
4:59 32°
Father Arrested In Death Of Brendan Creato
David Creato, Jr. Charged With Murder
CBSPhilly.com

(a)

creato david branden charged death murder father

All
News
Images
Videos
More
Settings
Tools

About 30,00,000 results (0.70 seconds)

Father of Brendan Creato indicted on murder charges | PhillyVoice
www.phillyvoice.com/father-brendan-creato-indicted-murder-charges/
Jan 11, 2016 - While full results of an autopsy on the body of three-year-old Brendan Link Creato are pending, preliminary results have been inconclusive regarding the boy's manner of death. ... The father of the 3-year-old toddler found dead last year in a Haddon Township park has been indicted on ...

Trial of Creato, charged with killing 3-year-old son, to start next week
www.philly.com/.../Trial-of-Creato-South-Jersey-man-charged-with-killing-3-year-ol...
Apr 12, 2017 - David "DJ" Creato, accused of killing his 3-year-old son Brendan, with ... the man charged in the case -- the boy's father, David "D.J." Creato Jr.

Murder trial begins for Westmont father - Courier-Post
www.courierpostonline.com/story/.../crime/.../creato-murder-trial...brendan/10065545...
Apr 19, 2017 - David "DJ" Creato is charged in connection with toddler son's ... Murder trial begins for Westmont father ... The jury took a trip to the site where 3-year-old Brendan Creato's body .... CAMDEN - A jury on Thursday began hearing the case of a Westmont man accused of killing his toddler son in October 2015.

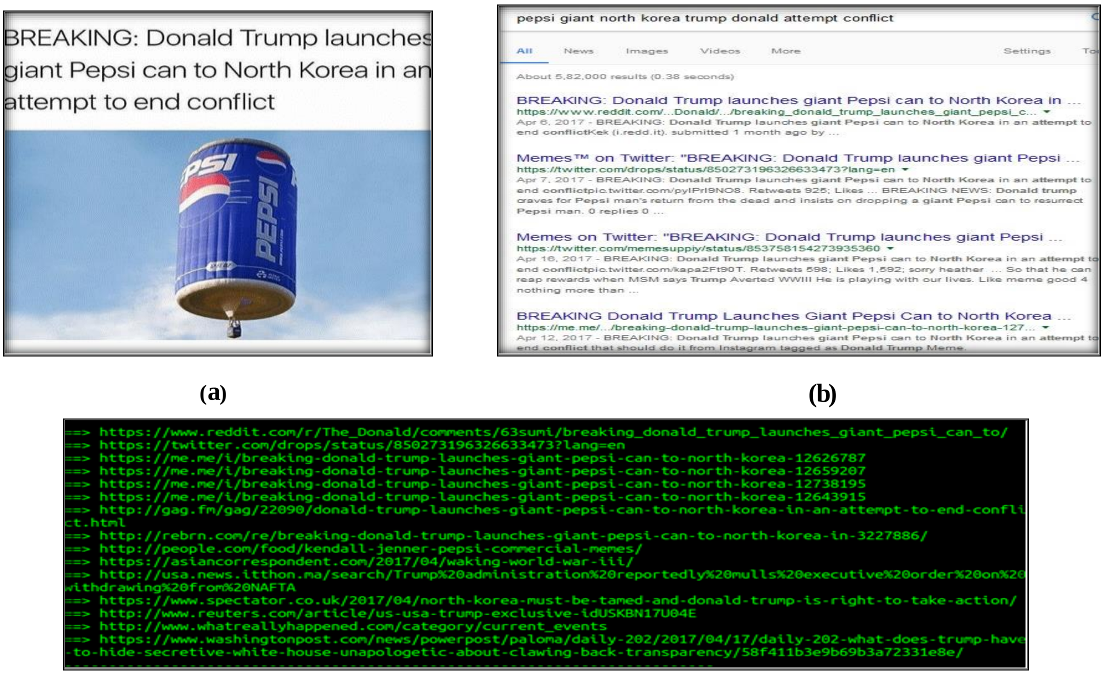
Prosecutor: Brendan Creato dad killed son to continue relationship ...
6abc.com/news/prosecutor-brendan-creato-dad-killed-son-to-continue.../1154775/
Jan 12, 2016 - May justice prevail," said D.J. Creato's father, David. EMBED More News ... D.J. Creato has been charged with murder for the death of his son.

(b)

```
==> http://www.phillyvoice.com/father-brendan-creato-indicted-murder-charges/
==> http://www.courierpostonline.com/story/news/crime/2017/04/19/creato-murder-trial-westmont-dj-brendan/100
655454/
==> http://www.cbsnews.com/news/david-creato-new-jersey-dad-charged-in-murder-of-son-brendan-creato/
==> http://www.fox29.com/news/72804042-story
==> http://abc7chicago.com/news/father-of-boy-found-dead-charged-with-murder/1154997/
==> http://www.philly.com/philly/news/new_jersey/creato-haddon-township-brendan-jersey-missing.html
==> https://patch.com/new-jersey/haddon/breaking-father-indicted-death-haddon-township-3-year-old
==> http://6abc.com/news/judge-delays-murder-trial-for-dad-of-brendan-creato/1537129/
==> http://www.nbcphiladelphia.com/news/local/DJ-Creato-Brendan-Creato-Father-Murder-Charges-365004801.html
==> http://www.nydailynews.com/news/crime/new-jersey-father-arrested-charged-murder-article-1.2493429
==> http://www.dailymail.co.uk/news/article-3394617/New-Jersey-dad-charged-murder-3-year-old-boys-death.html
==> http://www.phillymag.com/news/2016/01/11/david-creato-arrested-brendan-creato-death/
==> http://topics.nj.com/tag/brendan-creato/
==> http://www.nj.com/camden/index.ssf/2017/04/they_found_my_son_in_the_woods_mom_testifies_at_mu.html
==> http://abc7ny.com/news/prosecutor-new-jersey-dad-killed-3-year-old-son-to-continue-relationship/1157408/
```

(c)

Fig 3.1.4. Results on the real news (a) Input Image (b) Google search (c) Top 15 Google search


BREAKING: Donald Trump launches giant Pepsi can to North Korea in an attempt to end conflict
PEPSI

(a)

pepsi giant north korea trump donald attempt conflict

All
News
Images
Videos
More
Settings
To

About 5,82,000 results (0.38 seconds)

BREAKING: Donald Trump launches giant Pepsi can to North Korea in ...
https://www.reddit.com/...Donald/.../breaking_donald_trump_launches_giant_pepsi_c...
Apr 6, 2017 - BREAKING: Donald Trump launches giant Pepsi can to North Korea in an attempt to end conflictKek (i.redd.it). submitted 1 month ago by ...

Memes™ on Twitter: "BREAKING: Donald Trump launches giant Pepsi ...
https://twitter.com/drops/status/850273196326633473?lang=en
Apr 7, 2017 - BREAKING: Donald Trump launches giant Pepsi can to North Korea in an attempt to end conflictpic.twitter.com/pyIPrI9NO8. Retweets 925; Likes ... BREAKING NEWS: Donald trump craves for Pepsi man's return from the dead and insists on dropping a giant Pepsi can to resurrect Pepsi man. 0 replies 0 ...

Memes on Twitter: "BREAKING: Donald Trump launches giant Pepsi ...
https://twitter.com/memesuppiy/status/853758154273935360
Apr 16, 2017 - BREAKING: Donald Trump launches giant Pepsi can to North Korea in an attempt to end conflictpic.twitter.com/kapa2Ft90T. Retweets 598; Likes 1,592; sorry heather ... So that he can reap rewards when MSM says Trump Averted WWIII He is playing with our lives. Like meme good 4 nothing more than ...

BREAKING Donald Trump Launches Giant Pepsi Can to North Korea ...
https://me.me/.../breaking-donald-trump-launches-giant-pepsi-can-to-north-korea-127...
Apr 12, 2017 - BREAKING: Donald Trump launches giant Pepsi can to North Korea in an attempt to end conflict that should do it from Instagram tagged as Donald Trump Meme.

(b)

```
==> https://www.reddit.com/r/The_Donald/comments/63sumi/breaking_donald_trump_launches_giant_pepsi_can_to/
==> https://twitter.com/drops/status/850273196326633473?lang=en
==> https://me.me/i/breaking-donald-trump-launches-giant-pepsi-can-to-north-korea-12626787
==> https://me.me/i/breaking-donald-trump-launches-giant-pepsi-can-to-north-korea-12659207
==> https://me.me/i/breaking-donald-trump-launches-giant-pepsi-can-to-north-korea-12738195
==> https://me.me/i/breaking-donald-trump-launches-giant-pepsi-can-to-north-korea-12643915
==> http://gag.fm/gag/22090/donald-trump-launches-giant-pepsi-can-to-north-korea-in-an-attempt-to-end-confli
ct.html
==> http://rebrn.com/re/breaking-donald-trump-launches-giant-pepsi-can-to-north-korea-in-3227886/
==> http://people.com/food/kendall-jenner-pepsi-commercial-memes/
==> https://asiancorrespondent.com/2017/04/waking-world-war-iii/
==> http://usa.news.itthon.ma/search/Trump%20administration%20reportedly%20mulls%20executive%20order%20on%20
withdrawing%20from%20NAFTA
==> https://www.spectator.co.uk/2017/04/north-korea-must-be-tamed-and-donald-trump-is-right-to-take-action/
==> http://www.reuters.com/article/us-usa-trump-exclusive-idUSKBN17U04E
==> http://www.whatreallyhappened.com/category/current_events
==> https://www.washingtonpost.com/news/powerpost/paloma/daily-202/2017/04/17/daily-202-what-does-trump-have
-to-hide-secretive-white-house-unapologetic-about-clawing-back-transparency/58f411b3e9b69b3a72331e8e/
```

(c)

Fig 3.1.5. Results of a fake image after going through our system (a) Input Image (b) Google search (c) Top 15 Google search

Table 3.1.9, gives the variance in the value of reality parameter with the type of news. We observe that the system fails to classify the local news, as the value of Rp is very low for it. For national news, the value of Rp is quite effective (some false negatives will occur), whereas, for international news, the value is very effective. In this chapter, we have developed a novel algorithm that can detect fake news events. Reality parameter performed best when its value is 40%, which gives us 85% accuracy with the number of links being 15. Furthermore, the match between a summary of content and the search query seems to range between 0 to 48%. During our experimentation, a problem was faced during text extraction from images as for some images; we were not able to extract the text correctly because of various image characteristics like text with shadowing effect.

The proposed system addresses the fake news problem for both national and international news. The system seems to fail to classify local news, as the news does not get enough heat for major players to cover them.

Future work can be based on the improvement in the process of entity extraction for images text if the image is having a large amount of text, as this will directly affect the Google search results. Moreover, the integration of various social media handlers of credible media houses or newspapers for authentication of a news event with the current system might further improve the accuracy.

3.1.7 Significant Outcomes

The solution we have discussed in this section mainly relies on the build corpus of URLs for extracting crucial clues concerning the query and only those URLs are processed further that have a match in the build corpus. However, relying only on the build corpus restrict in fetching important clues as it may skip some URLs that may give some crucial information, that's why building corpus of authentic URLs is a challenging task. Furthermore, another thing is the proposed framework relies on content summarization techniques for analyzing the story of an article for the verification of content. Processing the complete article story is time-consuming as well as retrieving good summarization results is one of the challenging tasks. To overcome these challenges, we proposed another solution that relies on paragraph-based features instead of a complete article stories for retrieving the efficient clues. The detailed solution has been discussed in the following section. The following solution considered text as an input instead of text-embedded images.

3.2 Detecting FC: Text only part

The previous section discussed the detection of FC concerning to text-embedded images and to overcome the lack found in the previous method, another solution has been incorporated in this section. The solution is the enhanced version of the previous study, whereas in both the studies ultimately the text is given as an input.

Nowadays social media is one of the important mediums of sharing thoughts and opinions of the individual due to its easy access, but it also provides an opportunity for the malicious user to post deliberately fabricated false content to influence people for creating controversies, playing with public emotions, etc. The spread of contaminated information such as Rumours, Hoaxes, Accidental misinformation over the web is becoming an emergency that can have a very harmful impact on society and individuals. In this section, we discussed an automated system "Hoax-News Inspector" for the detection of fake news that propagates through the web and social media in the form of text. To distinguish fake and real reports on an early basis, we identified prominent features by exploring two sets of attributes that lead to information spread: article/post-content-based features, sentiment-based features, and the mixture of both called hybrid features. The proposed algorithm is trained and tested on the self-generated dataset as well as one of the popular existing datasets *Liar*. It has been found that the proposed algorithm gives the best results using the Random Forest classifier with an accuracy of 95% by considering all sets of features. Detecting and verifying news has many practical applications for business markets, news consumers, and time-sensitive services, which generally help to minimize the spread of false information. Our proposed system Hoax News-Inspector can automatically collect fabricated news data and classify it into binary classes Fake or Real, which later benefits further research for predicting and understanding Fake news.

Recent research encountered fake news in different ways. One of the methods is via crawling the web and analyzing the article content. In [77], the author focuses on the headline of fake and real news; from the study, it has been found that fake news has a significant difference compared to real news. Whereas the authors of [46], applied a warning system on news articles, that can help users to better understand the credibility of the news article. In [84], authors crawl and analyze the google search results and applied a rule-based approach to predict whether the news is fake or not. These approaches are leading to some promising results. However, no principle study is conducted on paragraph features of an article. From the analysis, it has been found that each paragraph leads to giving important features (content and sentiment-based features) for prediction. Besides, there has been no research that provides a standardized understanding of (i) what possible paragraph features in a news article are; (ii) how discriminative these features are. To give a comprehensive understanding of these aspects, here we made the following key contributions that are summarized below:

- A classification framework is developed that uses a content similarity-based agreement approach on web URLs for authenticating the credibility of content, to detect fake news on an early basis.
- A new methodology is outlined that leverages a rumor debunking system (Snopes, Politifact) and Wiki database to build a dataset for early detection of fake news on the web and social media.
- The proposed framework can handle long articles for analysis, and in most of the real scenarios, articles are long and descriptive, to have a good representation of an event.
- The performance of the proposed approach is evaluated on public and self-generated datasets of Fake News, and also it is compared with the existing state-of-the-art.

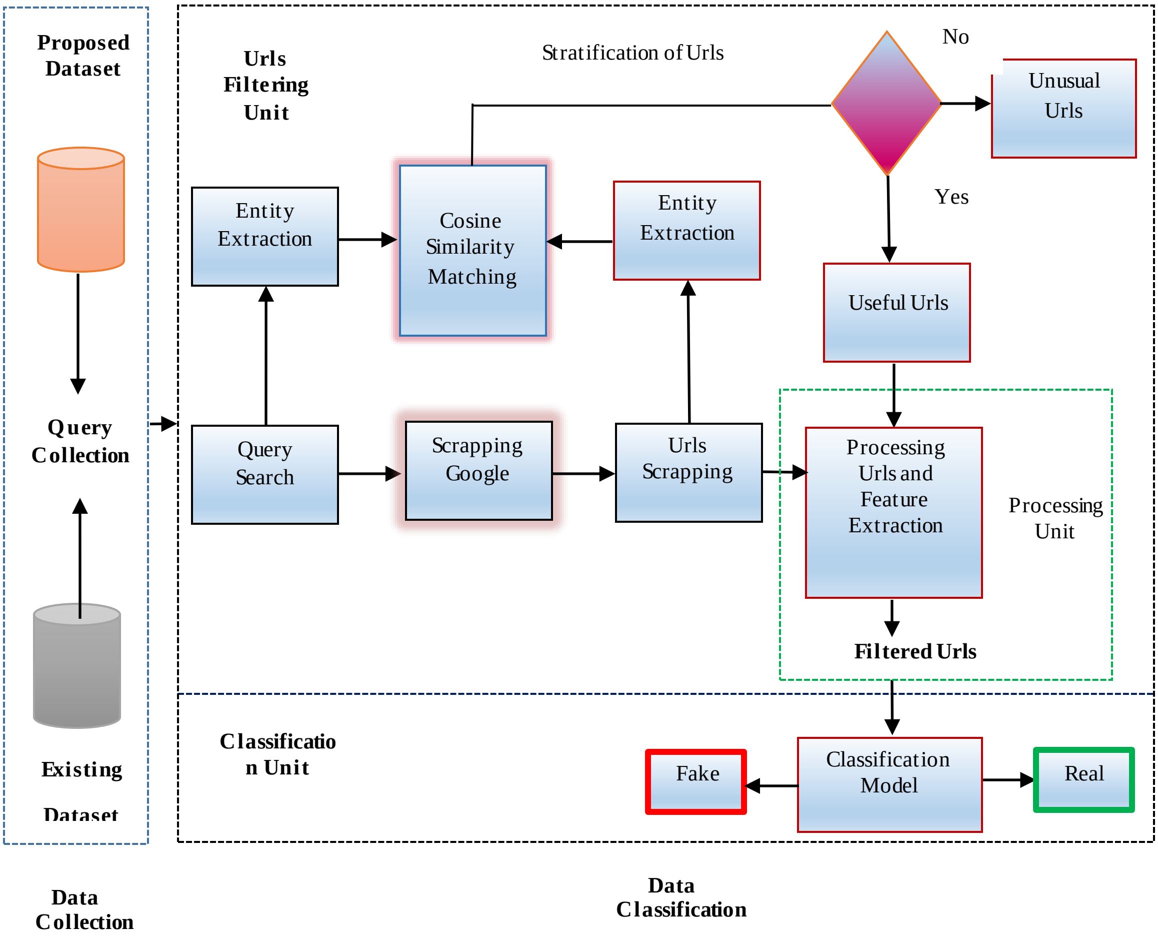


Fig.3.2.1. Proposed Architecture

3.2.1 Data Collection

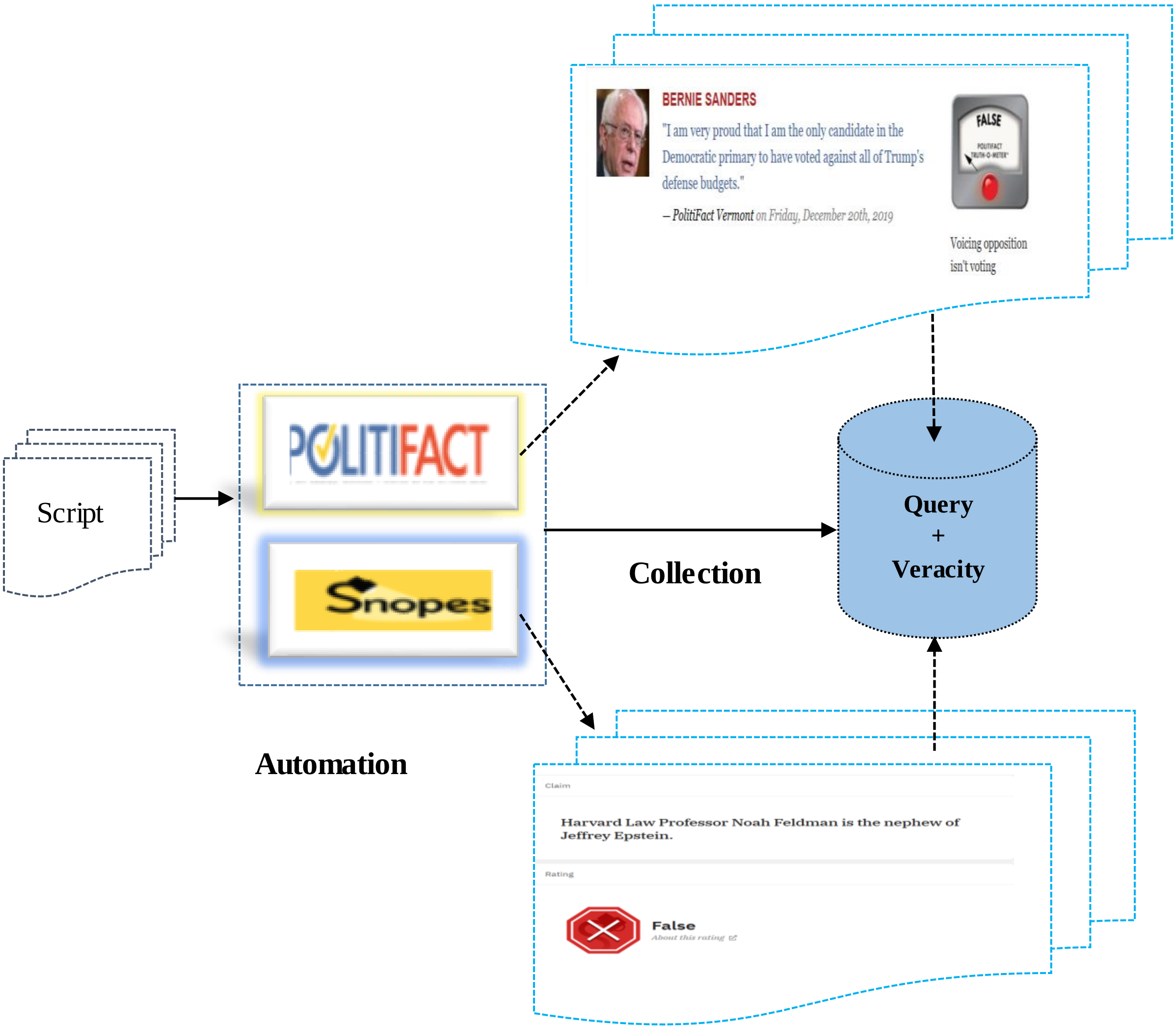
The basic overview of the proposed idea is shown in Fig.3.2.1, where the system is segregated into two units data collection and data classification. The data collection unit is responsible for collecting samples from rumor debunking websites and the Wiki database using the proposed script concerning the posted claims related to different subjects like politics, death reports, etc. The collected dataset will further be used to retrieve features from credible web sources and later utilized for the classification of claims in real or fake. It can be observed from Fig.3.2.1 that the proposed algorithm is a build-up of small modules, that are combined to form an overall system. The main goal of the algorithm is to detect whether the post is fake or not. The data classification unit is further subcategorized into modules as i) URLs Filtering Unit ii) Processing Unit iii) Classification Unit. These modules are deeply discussed in the following section

The first module in the series of Fake news detection is the Data Collection. The collection of fake or real data is one of the challenging tasks. Many of the authors have used prominent fact-checking websites as one of the sources for the collection of claims; some of them are Politifact, Snopes, Fact check, etc. Here, we are mainly focused on the Politifact and Snopes websites for collecting the claims and their veracity for building ground truth data. The data collection module is responsible for collecting ground truth data that further be used to build the model. Two different aspects are employed in the data collection process. Firstly, we have used some popular existing datasets, one of them is “Liar”. Liar[17] dataset is one of the publicly available datasets that has been used by [85].

The dataset contains 12.8K short statements from Politifact.com that comprise six labels of truthfulness rating: pants-fire, false, barely-true, half-true, mostly-true, and true. In our work, we mainly focused on categorizing information into two categories fake and real. So, for binary classification of news like [85], the same concept has also been applied in our work concerning fake and real samples(pants-fire, false, barely-true are considered as fake and half-true, mostly-true and true are as real). This dataset is comprised of a significant amount of posts from online social media that mostly deal with political issues, including statements of democrats and republicans. Except for the existing datasets, we have also collected our dataset (Primary dataset) from rumor debunking websites (Politifact and Snopes). The proposed dataset is the collection of 671 claims related to different categories (politics, death…etc.) containing 255

[17] https://www.cs.ucsb.edu/~william/data/liar_dataset.zip

real claims and 416 false claims. The technical view of the data collection process for our dataset is shown in Fig.3.2.2, whereas, Algorithm 3.2.1 shows the process we have adopted for the data collection of our proposed dataset. In this algorithm, we have discussed in detail how claims and respective veracity are extracted from debunking websites (Politifact and Snopes). Selenium web driver has been used to automate the system of collection, where firstly the web driver retrieves the page where the claim has to be extracted. Here n is the number of pages through which we want to collect the claims. By specifying the given XPath, we can extract specific attributes (Claim and Veracity) from the page. This set of collected claims as a query is the input for the next unit.



BERNIE SANDERS

"I am very proud that I am the only candidate in the Democratic primary to have voted against all of Trump's defense budgets."

– PolitiFact Vermont on Friday, December 20th, 2019

Voicing opposition isn't voting

Claim

Harvard Law Professor Noah Feldman is the nephew of Jeffrey Epstein.

Rating

False

About this rating

Fig.3.2.2. Data collection process for the Proposed Data set

Algorithm 3.2.1 Data Collection

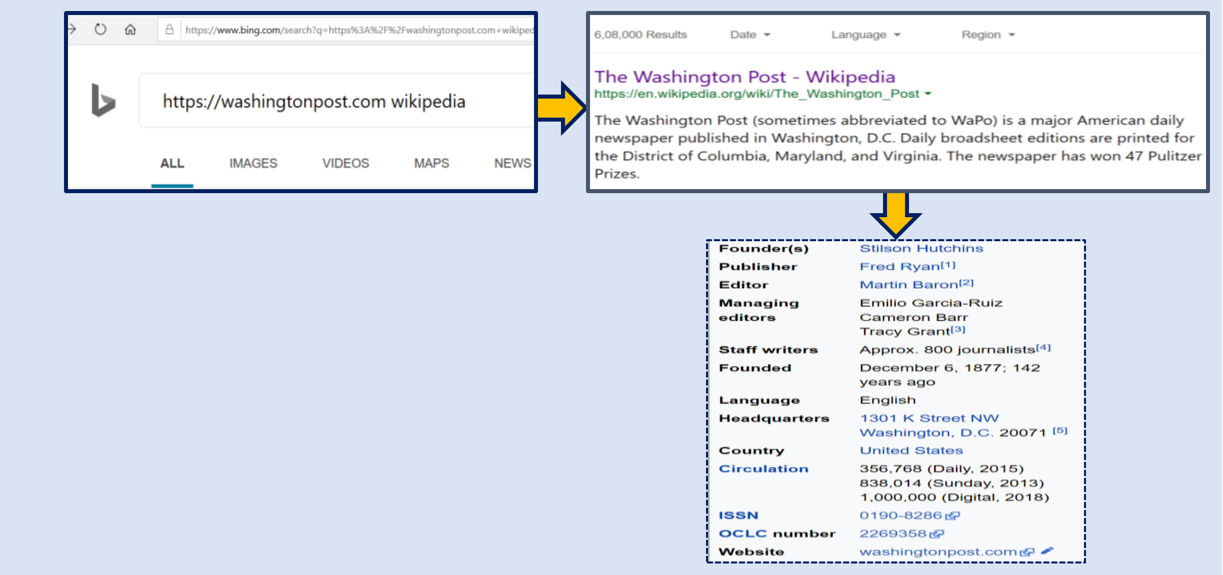
```
procedure data collection ()
driver = web driver. Chrome ()
driver.get('https://www.politifact.com/truth-o-meter/statements/')
pages= n
For k in range (n, page+1):
        Scrap(Claims)
        Scrap(Veracity)
        end procedure
```

3.2.2 Data Classification

The second module in the series of fake news detection is the data classification. This module is further subdivided into three submodule URLs Filtering, Processing Unit, and Classification unit. This section is covering the technical description of all three modules in detail. The URL filtering unit is the first unit of data classification and is responsible for filtering URLs that are needful, later in the second phase the set of useful URLs moves to the processing and feature extraction unit for retrieving efficient clues required for the prediction of fake news from these URLs. Lastly, all sets of features are fed into the classification model for predicting results. The detailed description of each of these phases is shown in the following sections.

- **URLs Filtering Unit**

This is the first submodule of the data classification. This module is responsible for filtering useful URLs coming out from the google search results. In the first phase, the query which is


https://www.bing.com/search?q=https%3A%2F%2Fwashingtonpost.com+wikiped
ALL
IMAGES
VIDEOS
MAPS
NEWS
6,08,000 Results
Date
Language
Region

https://washingtonpost.com wikipedia

The Washington Post - Wikipedia
https://en.wikipedia.org/wiki/The_Washington_Post

The Washington Post (sometimes abbreviated to WaPo) is a major American daily newspaper published in Washington, D.C. Daily broadsheet editions are printed for the District of Columbia, Maryland, and Virginia. The newspaper has won 47 Pulitzer Prizes.

Founder(s)	Stilson Hutchins
Publisher	Fred Ryan[1]
Editor	Martin Baron[2]
Managing editors	Emilio Garcia-Ruiz Cameron Barr Tracy Grant[3]
Staff writers	Approx. 800 journalists[4]
Founded	December 6, 1877; 142 years ago
Language	English
Headquarters	1301 K Street NW Washington, D.C. 20071 [5]
Country	United States
Circulation	356,768 (Daily, 2015) 838,014 (Sunday, 2013) 1,000,000 (Digital, 2018)
ISSN	0190-8286
OCLC number	2269358
Website	washingtonpost.com

Fig 3.2.3. Process of verifying the source of Useful URLs

coming out from the previous module is given as an input, which further is used to extract google search results. In the next step for each query, entities have been extracted. Rake library is used to extract the entities. Rake[18] is termed as Rapid Automatic Keyword Extraction is one of the well-known and widely used NLP techniques to automatically extract keywords from sentences. Entity extraction is responsible for providing a compact representation of content.

As we would like to process only those URLs that relate to our query. For retrieving useful URLs, a Cosine Similarity algorithm has been applied between entities extracted from the search query and the entities extracted from each URL's title, retrieved in the google response results. The filtering of URLs is an important unit as this will give the set of reliable data that we need to process further. To decide the rule, which URL has to process, the cosine similarity score of the retrieved titles in the google search results is analyzed thoroughly. If a match is above the threshold value and the particular keyword is found in the title, then the URL is considered as a Useful URL, and these sets of URLs further be given as an input to the next unit. For specifying the set of keywords, we have analyzed the news titles and observed that there exists a pattern of specific keywords in the titles encountering false content. One of the rumors debunking website[19] has been analyzed thoroughly and manually extracted the list of ['false', 'fake news, 'falsely', 'mistakenly', 'incorrectly', 'misinformation', 'death hoax', 'false news', 'hoaxes', 'hoax', 'rumor', 'rumour'] frequently used words in the titles encountering fake news and included them in our keyword list. The complete procedure of filtering URLs is described in Algorithm 3.2.2. The first decision of the threshold for the stratification of URLs is based on the manual analysis of all the samples value in the feature set. We have manually identified the relevant titles concerning the search claim and for all those relevant titles retrieved, we combinedly take an average value that comes out to be 0.31 and use it as a threshold value for the initial stratification of URLs. It also reveals that the relevant links have a cosine similarity value range that lies above 0.30. Hence, if the cosine similarity value is found to be greater than 0.30 and the specific keyword is encountered as mentioned in the keyword list or the cosine similarity value is greater than 0.30 the corresponding Url is returned otherwise not considered for analysis. After this phase of filtering, another phase of filtering is applied by removing URLs not are from authentic sources. Instead of manually building a dictionary of the authentic news channel and making it restricted, we applied the dynamic verification of URLs. The respective URL has been searched over google, and their Wikipedia

[18] https://pypi.org/project/rake-nltk/
[19] https://www.snopes.com/fact-check/celebrity-death-hoaxes/

source has been crawled for retrieving useful attributes like Type, Publisher, Owner, Founded, Country, Circulation. These attributes are responsible for identifying the reliability of the respective URLs, as shown in Fig 3.2.3. In the last phase of URLs Filtering, we manually analyzed all the collected attributes and removed any sort of irrelevant data. The final set of filtered titles are given as an input to the next unit.

Algorithm 3.2.2 URLs Filtering

```
procedure URLs Filtering (title, query, keywords)
CS ← Cosine Similarity (query, title)
If CS>0.30 and the title has "keywords" or CS>0.30
    return URLs
Else
    return false
end procedure
```

- **Processing Unit and Feature Extraction**

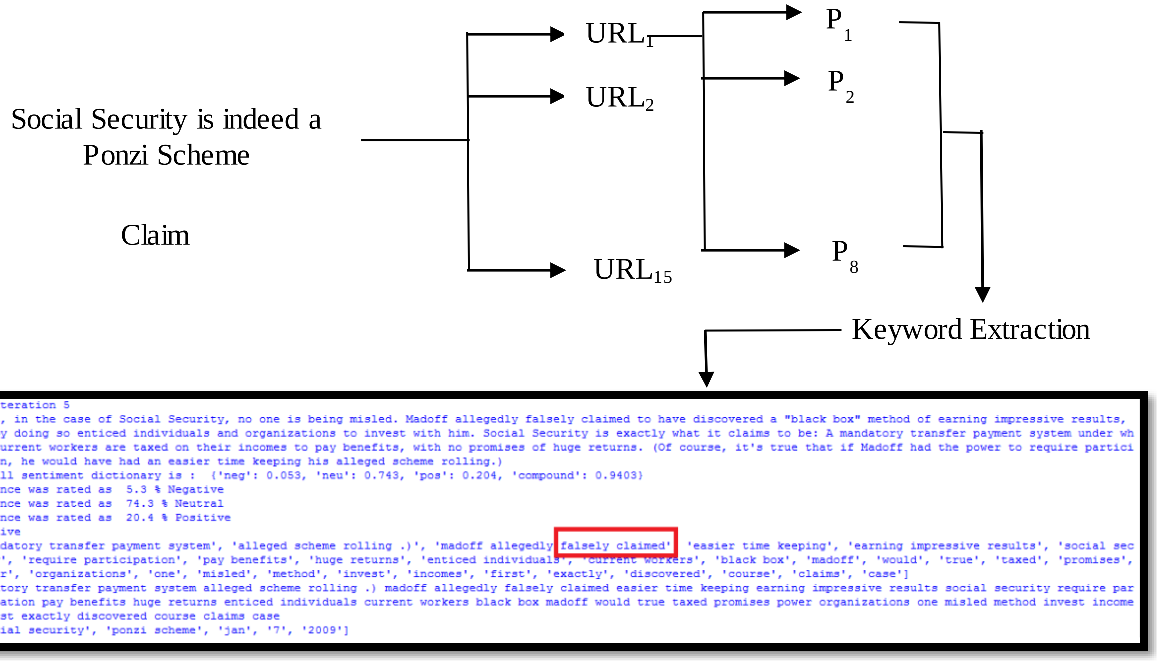
This is the second submodule of the data classification. This module is responsible for the processing of useful URLs that are given as input from the previous phase. In this phase, the content of the useful URLs are processed. As we have discussed earlier that we are going to consider long articles. So for processing, articles having at least 8 paragraphs are considered for analysis, it has been observed from [20] that on average the length of good articles lies within 4 to 8 paragraphs, so we have chosen the max value for analysis. The first 8 paragraphs of each title are extracted and processed. A beautiful soup library is used for the extraction of paragraphs from the webpage, and the Rake library of python is utilized for the entity extraction. One of the examples of encountered keywords from the extracted entities having a false claim is shown in Fig.3.2.4. In this unit, another set of filtering is applied. The filtering is relying on the average cosine similarity match value between the entities extracted from the title and the entities extracted from each paragraph. If the Average Total Cosine Similarity Score(ATCS)>=0.45, then the URL is considered to be the useful URL, otherwise not. The Average Total Cosine Similarity Score (ATCS)>=0.45 is calculated by applying an experiment on the dataset considered for analysis. The threshold value of 0.45 is decided based on analyzing the ATCS value concerning all fake and real samples. To decide the threshold, the average value of ATCS for all the fake and real samples has been analyzed, where it has been found that the average value of ATCS concerning the total number of samples in the feature set is come out to be below 0.45. So, in this way, we have considered the threshold value to be ATCS>=0.45 to focus on the relevant set of URLs for the reliable and accurate prediction that

[20] https://ezinearticles.com/?Article-Writing---How-Many-Paragraphs-Should-My-Article-Be?&id=4019539

further goes to the classification unit. The entire process of this phase is explained in the following steps and Fig.3.2.5. In the first step titles coming from the previous unit are further be processed and filtered concerning the article length. The title will only be processed also if the article associated with the title have at least 8 paragraphs. In the second step, the first 8 paragraphs have been extracted and processed for each given title. For each given paragraph concerning the given title, entities are extracted, and features have been analyzed. In the third step, to find the similarity between the title concerning each paragraph in the article, the cosine similarity score is calculated. The total cosine similarity score can be calculated by summing up the cosine similarity for every 8 paragraphs; later the average value is taken into consideration for the 8 paragraphs, represented in the Algorithm 3.2.3 termed as ACS (Average of Total Cosine similarity of all 8 paragraphs). In the last step to decide the threshold value for further stratification of titles with respect to the content of the article, the AVG function is applied over an attribute ACS considering all samples in the feature set concerning this attribute that further is used to calculate ATCS (Average of Total Cosine Similarity) that comes out to be 0.45. So, when ATCS<0.45, URL is not considered for analysis. The filtered set of titles is considered as useful titles and goes for the classification in the next unit for reliable prediction.

- **Classification Unit**

The last submodule of this system is the classification unit. In this unit, the final set of filtered URLs is going to process further and is considered useful URLs. The extracted features with respect to each useful URL, as discussed in the processing unit have been used to build the



```
The iteration 5
First, in the case of Social Security, no one is being misled. Madoff allegedly falsely claimed to have discovered a "black box" method of earning impressive results,
and by doing so enticed individuals and organizations to invest with him. Social Security is exactly what it claims to be: A mandatory transfer payment system under wh
ich current workers are taxed on their incomes to pay benefits, with no promises of huge returns. (Of course, it's true that if Madoff had the power to require partici
pation, he would have had an easier time keeping his alleged scheme rolling.)
Overall sentiment dictionary is :  {'neg': 0.053, 'neu': 0.743, 'pos': 0.204, 'compound': 0.9403}
sentence was rated as  5.3 % Negative
sentence was rated as  74.3 % Neutral
sentence was rated as  20.4 % Positive
Positive
['mandatory transfer payment system', 'alleged scheme rolling .)', 'madoff allegedly falsely claimed'  'easier time keeping', 'earning impressive results', 'social sec
urity', 'require participation', 'pay benefits', 'huge returns', 'enticed individuals', 'current workers', 'black box', 'madoff', 'would', 'true', 'taxed', 'promises',
'power', 'organizations', 'one', 'misled', 'method', 'invest', 'incomes', 'first', 'exactly', 'discovered', 'course', 'claims', 'case']
mandatory transfer payment system alleged scheme rolling .) madoff allegedly falsely claimed easier time keeping earning impressive results social security require par
ticipation pay benefits huge returns enticed individuals current workers black box madoff would true taxed promises power organizations one misled method invest income
s first exactly discovered course claims case
['social security', 'ponzi scheme', 'jan', '7', '2009']
```

Fig. 3.2.4. Example of encountered false keyword from the extracted entities having false claim

classification model and to classify whether a post is fake or real. Table 3.2.1 describes all sets of features we have utilized for analysis.

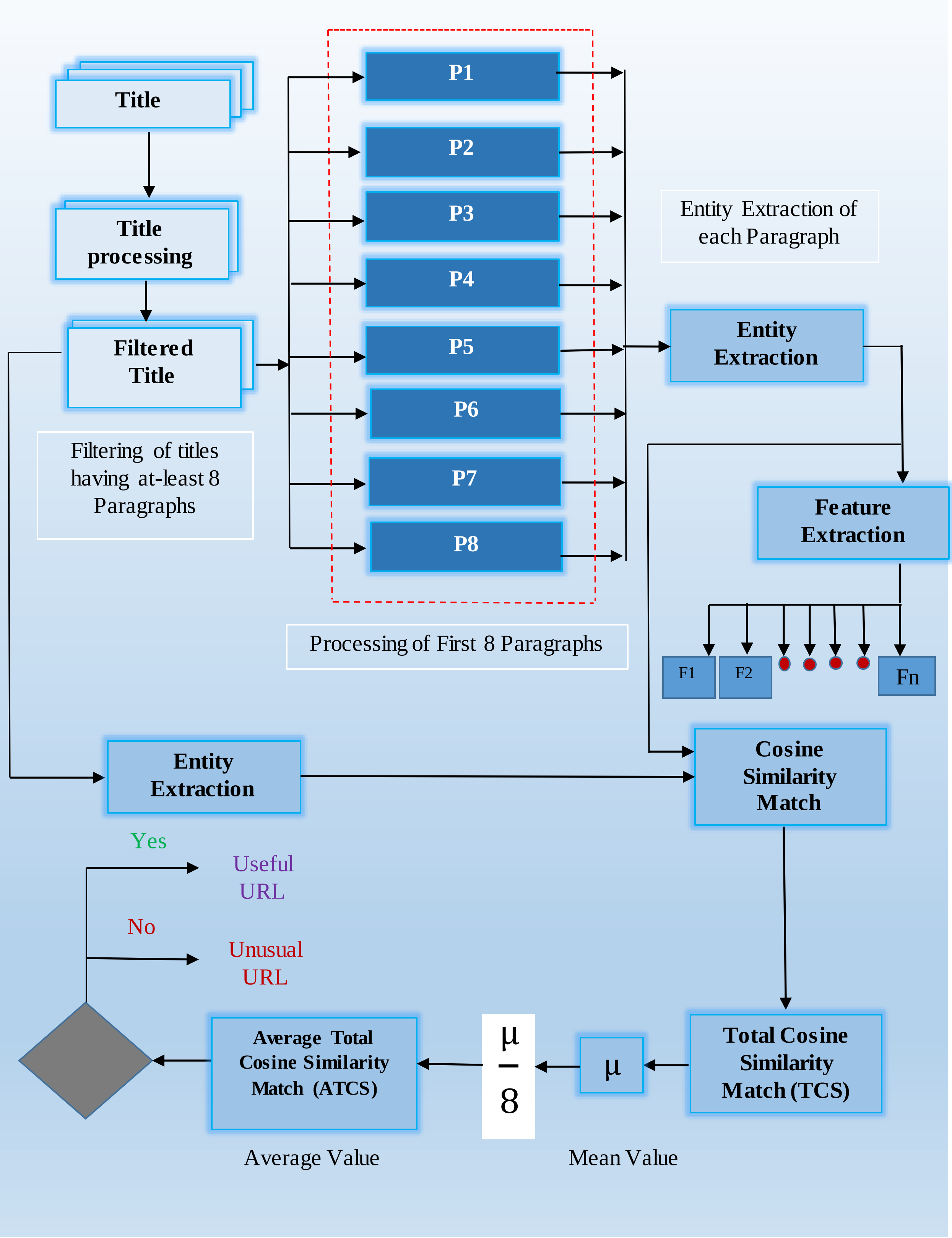

Title
Title processing
Filtered Title
Filtering of titles having at-least 8 Paragraphs
P1
P2
P3
P4
P5
P6
P7
P8
Processing of First 8 Paragraphs
Entity Extraction of each Paragraph
Entity Extraction
Feature Extraction
F1
F2
Fn
Cosine Similarity Match
Entity Extraction
Yes
Useful URL
No
Unusual URL
Average Total Cosine Similarity Match (ATCS)
μ / 8
μ
Total Cosine Similarity Match (TCS)
Average Value
Mean Value

Fig.3.2.5. Flow Diagram of Processing Unit

Table 3.2.1. Proposed Features

Feature Type	Feature	Feature Description
Content-Based	Fake_count	The total number of fake keywords in each paragraph text of an article.
Content-Based	Min_fake_count	Min count of the fake keyword in each paragraph of an article.
Content-Based	Max_fake_count	Max count of the fake word in each paragraph of an article.
Content-Based	TS	Title Cosine similarity score with the given query.
Content-Based	TCS	The total cosine similarity score between title heading and each paragraph text.
Content-Based	ACS	Average of Total Cosine similarity Score of all 8 paragraphs.
Sentiment Based	TCPS	Total Count of Positive Sentiments in all paragraphs
Sentiments Based	TCNES	Total Count of Negative Sentiments in all paragraphs.
Sentiment Based	PTN	Positive to Negative ratio
Sentiment Based	NTP	Negative to Positive ratio
Sentiment Based	TCNS	Total Count of Neutral Sentiments

Algorithm 3.2.3. Processing Unit

```
procedure Processing Unit (URLs)
for k in range (1,9):
    if(soup.find_all("p")  [k].get_text ()):
        paragraph=soup. find_all("p")  [k]. get_text ()
        $CS_1 \leftarrow$ Cosine_Similarity  (paragraph, title)
        simi_score_total= simi_score_total+$CS_1$
        Feature_extraction(paragraphs)
ACS= ((simi_score_total)/8)
Average_Total_Cosine Similarity  (ATCS)= AVG(ACS)
If  ATCS>=0.45
    return URL
Else
    return False
end procedure
```

3.2.3 Experiment and Results

In this section, we first illustrate the experimental setup and the baselines. Then, we showed the observed results and compared various models. The experiment has been performed on two sets of data. The first one is on the publicly available dataset for fake news analysis: Liar dataset and another one on the self-generated dataset. We have four baselines and among four baselines majority baseline, are logistic regression classifier(LR), a support vector machine(SVM). In this section, we provide an in-depth performance analysis of our traditional machine learning classifiers. We have also highlighted the best performance for each dataset and each matrix in bold. The evaluation measures that have been taken for analysis are accuracy, tp-rate, precision, recall, and f1-score to distinguish news, fake, and real.

- **Liar Dataset**

Liar dataset was created by [57], which is the collection of 12,836 short-statements along with its veracity. The dataset is labeled with a discrete set of values from 1 to 6 corresponding to pants-fire, false, mostly-false, half-true, mostly-true, and true. As our problem is based on binary class classification so the data with the first, three labels are labeled as fake news, and the others are labeled as real news. The proposed solution has been applied to the given dataset, and the detailed evaluation measure like Accuracy, TP-rate, FP-rate, Precision, Recall, F-Measure has been shown in Table 3.2.2.

Table 3.2.2. Performance of the Model on the Liar dataset.

Feature	Classifier	Accuracy	TP-Rate	FP-Rate	Precision	Recall	F-Measure
Content based	**LR**	0.79	0.79	0.21	0.84	0.79	0.78
Content based	**Random Forest**	**0.95**	0.95	0.04	0.96	0.95	0.95
Content based	**Naïve- Bayes**	0.78	0.78	0.21	0.82	0.78	0.77
Content-based	**Decision Tree**	0.81	0.81	0.18	0.85	0.81	0.80
Content based	**k-NN**	**0.94**	0.94	0.05	0.94	0.94	0.94
Content based	**SVM**	0.63	0.63	0.36	0.78	0.63	0.57
Sentiment based	**LR**	**0.83**	0.83	0.16	0.83	0.83	0.83
Sentiment based	**Random Forest**	0.86	0.86	0.13	0.87	0.86	0.86
Sentiment based	**Naïve-Bayes**	**0.83**	0.83	0.16	0.83	0.83	0.83
Sentiment based	**Decision Tree**	**0.87**	0.87	0.12	0.89	0.87	0.86
Sentiment based	**k-NN**	0.82	0.82	0.17	0.83	0.82	0.82
Sentiment based	**SVM**	0.53	0.53	0.46	0.54	0.53	0.51
Hybrid feature (Content+sentiment)	**LR**	0.79	0.79	0.20	0.85	0.79	0.78
Hybrid feature (Content+sentiment)	**Random-Forest**	0.93	0.93	0.06	0.93	0.93	0.93
Hybrid feature (Content+sentiment)	**Naïve-Bayes**	0.78	0.78	0.21	0.83	0.78	0.77
Hybrid feature (Content+sentiment)	**Decision-Tree**	0.81	0.81	0.18	0.85	0.81	0.80
Hybrid feature (Content+Sentiment)	**k-NN**	0.89	0.89	0.10	0.89	0.89	0.89
Hybrid feature (Content+sentiment)	**SVM[RBF]**	**0.64**	0.64	0.35	0.79	0.64	0.59

From Table 3.2.2, it has been observed that the evaluation is performed using various machine learning classifiers (Logistic Regression (LR), Random-Forest, Naïve Bayes, K-NN, and SVM) concerning individual proposed features. Weka tool is used to evaluate the performance of the classifier. Weka is one of the open-source machine learning software that gives transparent access to well-known toolboxes such as sci-kit-learn, R, and Deep-learning. From Table 3.2.2, it is depicted that the Random-Forest classifier outperforms all others. From the previous studies also, it has been seen that Random-Forest and SVM were found to be the best classifier for detecting fake news. The analysis has been performed for the given classifier with respect to each type of feature (Content-based, Sentiment-based, and Hybrid-based) exploited for fake news detection. It can be seen that random forest outperforms others in each of the given types of feature, having an accuracy of 95% while considering only content-based features, 86% when considering sentiment-based features, and 93% while considering both the features together. Naive Bayes and Decision Tree classifiers perform best when considering only Sentiment-based features, while K-NN gives an accuracy of 94% when considering only content-based features. The comparison analysis has also been performed with the existing literature on the Liar dataset. The detailed evaluation measure, like Accuracy, Precision, Recall, and F-Measure is considered for analysis, as shown in Table 3.2.3.

Table 3.2.3. Prediction comparison with the existing literature on Liar dataset

Reference	Model	Accuracy	Precision	Recall	F1score
[85]	**SVM**	0.56	0.57	0.56	0.48
	LR	0.56	0.56	0.56	0.51
	Decision-Tree	0.51	0.51	0.51	0.51
	Ad boost	0.56	0.56	0.56	0.54
	Naïve-Bayes	0.60	0.59	0.60	0.59
	k-NN	0.53	0.53	0.53	0.53
[57]	**SVM**	0.25	-	-	-
	LR	0.24	-	-	-
[86]	**SVM**	0.57	0.59	0.53	0.54
[87]	**SVM**	0.73	0.64	0.70	0.67
Our Method	**Random Forest**	**0.95**	0.96	0.95	0.95
	LR	0.83	0.83	0.83	0.83
	Naïve Bayes	0.83	0.83	0.83	0.83
	Decision Tree	0.87	0.89	0.87	0.86
	k-NN	0.94	0.94	0.94	0.94
	SVM[RBF]	0.64	0.79	0.64	0.59
	SVM[Linear]	0.93	0.94	0.93	0.93

From the given Table 3.2.3, it can be observed that the proposed method outperforms other methods [85][57][86][87] applied to the same dataset. Here also, it can be observed that the Random Forest classifier outperforms all others with an accuracy of 95%. Most of the authors have applied the SVM classifier for evaluation, as we already discussed previously. It has been observed that among all given methods our method performs best on SVM, with an accuracy of 93% when applying Linear kernel, while [57] performs worst using an SVM classifier.

- **Proposed Dataset**

Fake news prediction is one of the new fields, due to which very few standard datasets are available for evaluation. So, we have also included one of the key contributions in this study by building our dataset and testing the proposed architecture on it. The dataset is the collection of 671(416 False statements and 255 True statements) short statements along with its veracity, collected from two well-known rumor debunking websites “Politifact” and “Snopes” related to the topics of politics and death hoaxes, along with it we have also used wiki death reports data[21]. The selenium Web driver has been used to automate the rumor debunking website for the collection of facts and veracity (True/False). The performance analysis of the model on the proposed dataset is described in Table 3.2.4.

Table 3.2.4. Performance of the Model on the proposed dataset

Feature	Classifier	Accuracy	TP-Rate	FP-Rate	Precision	Recall	F-Measure
Content based	**LR**	**0.96**	0.96	0.039	0.962	0.96	0.96
Content based	**Random Forest**	**0.95**	0.95	0.04	0.95	0.95	0.95
Content based	**Naïve-Bayes**	0.68	0.68	0.31	0.74	0.68	0.66
Content based	**k-NN**	**0.95**	0.95	0.04	0.95	0.95	0.95
Content based	**SVM**	0.59	0.59	0.39	0.66	0.59	0.55
Sentiment based	**LR**	0.95	0.95	0.04	0.95	0.95	0.95
Sentiment based	**Random-Forest**	**0.95**	0.95	0.04	0.95	0.95	0.95
Sentiment based	**Naïve-Bayes**	**0.94**	0.96	0.03	0.96	0.96	0.96
Sentiment based	**k-NN**	0.94	0.94	0.05	0.94	0.94	0.94
Sentiment based	**SVM**	0.56	0.56	0.43	0.56	0.56	0.56
Hybrid feature(Content+sentiment)	**LR**	0.93	0.93	0.06	0.93	0.93	0.93
Hybrid feature(Content+sentiment)	**Random-Forest**	**0.94**	0.94	0.05	0.94	0.94	0.94
Hybrid feature(Content+sentiment)	**Naïve-Bayes**	0.68	0.68	0.30	0.74	0.68	0.66
Hybrid feature(Content+sentiment)	**k-NN**	0.93	0.93	0.07	0.93	0.93	0.93

Logistic regression is also found to be one of the good classification models in the previous studies, for the detection of false information [88]. From the above table, it has been observed that the Logistic Regression classifier outperforms others with an accuracy of 96% by considering only content-based features, whereas Random forest gives the best results with an accuracy of 95% by considering only sentiment-based features, while it has been found that k-NN gives the best accuracy by considering only content-based features instead of using sentiment and hybrid-based features.

3.2.4 Significant Outcome

In this section, we provide a model called "Hoax News Inspector" which provides general solutions for data collection and data classification towards fake news detection. A data collection strategy is provided to collect data, and a machine learning solution is employed to detect fake news. We have used a content-based, sentiment-based, and hybrid feature for the detection task. It has been observed that random forest classifier outperforms others in each of

[21] https://en.wikipedia.org/wiki/Wikipedia:Database_reports/Recent_deaths

the given types of features, with an accuracy of 95%. Fake news detection is one of the emerging topics, and there are several interesting options for future work. One of the important works that can be done is to build an image dataset that is embedded with textual data as very few datasets are available related to this. Also, our proposed framework could be extended to detect false information in real-time by implementing the extension. In the future, we are extending this work with respect to multimedia as well. The work can be extended with respect to multimedia data where a voice in a video can be processed via translating the speech content using google speech API into text and applying the analysis over it. We can extend this work with an image as well via extracting the text claim embedded in an image using an OCR technique, and extracted text can be processed that further be utilized in the verification of images attached with a false claim.

Chapter 4

Detection of Fraudulent Claims Accompanying Multimedia Data

This chapter incorporates two different methods to detect fraudulent/deceptive claims accompanying multimedia content on web platforms. The two different methods included the following cases. The first case is when the text/claim accompanies a video, the second case included the text/claim accompanying with an image. In the thesis, the methodologies concerning each of the given cases have been discussed in detail. Further, the classification results of the proposed approaches are validated on standard datasets and compared with existing state-of-the-art methods.

In the previous chapter, we have discussed the techniques considering only text query as an input, and no other media content is attached to justify the claim. Whereas, in this chapter we considered the claim accompanying multimedia content (images and videos). Firstly, we discussed the technique where the claim is accompanying with image content and secondly we discussed the technique where the claim is accompanying with video content.

4.1 Detecting fraudulent content accompanying image

The verification of multimedia content over social media is one of the challenging and crucial issues in the current scenario and gaining prominence in an age where user-generated content and online social web platforms are the leading sources in shaping and propagating news stories. These sources allow the user to share their opinion without any restriction and opportunistic users are often utilized for posting misleading content on social media such as Twitter, Facebook, etc. In the current scenario to lure users towards the news story, the text is often attached with some multimedia content(images/videos/audios). Verification of this content to maintain the credibility and reliability of the information on social media is paramount important. Motivated by this, here we present a generalized system that supports the automatic classification of images into credible or misleading. A detailed description of the proposed framework has been discussed in the following paragraphs.

4.1.1 Problem Description

In this section, we describe the problem description and briefly explain the generalized model for the verification of multimedia content posted on social media. The multimedia post we have considered here is incorporating two parts 1. Image Part 2. Claim Part. Any post that is associated with these two parts is called as a *multimedia post*. We define two sets of crucial features (Content-based and Semantics-based). The efficient clues that have been retrieved for the prediction of misleading content are from two parts (1) Tweet + Image, (2) Image only. The first part incorporates both tweets and images for the retrieval of efficient clues. The important clues considered in this category are based on semantic similarity, trace of fake, and trace of doubt. Along with this, the hidden representation of word sequences has been generated using the Bi-directional LSTM model. Whereas, the second part (Image only) of analysis has been applied by extracting crucial knowledge from the existing instances of an image on web. The feature-based evidence with respect to each multimedia post can be represented as $m^i = (DB^i, UNS^i, S^i)$ Where, DB^i defines that the user is in doubt with the claim and with the accompanying multimedia item(image) from an event. The UNS^i defines that the user is not in support of the claim and not confident with the accompanying multimedia item(image) from an event. Whereas, S^i defines the semantic similarity score. These crucial factors are identified with respect to each multimedia post for the prediction of misinformation. Each multimedia post is associated with 'k' claims posted by 'r' users. Multiple users share their different opinion with respect to an individual image. The aim is to verify the given claim and the accompanying multimedia item(image) from an event are faithfully describing each other and not contradictory. It further returning a binary decision representing verification of whether the multimedia item reflects the reality of the event in the way purported by the tweet. The image traces have been analysed over the top 10 returned web responses to gather crucial evidence for the prediction of misleading information. From the empirical analysis, it has been found that for some of the images relevant claims are not retrieved. The google search engine is not able to identify the images in the correct context, and due to which useful search results may not be retrieved. To resolve such a scenario, we have also retrieved traces of an image from Microsoft Bing visual search[22]. Some of the results responses from Microsoft visual search and Google image search[23] concerning to an image has been shown in Table 4.1.2 and the results reveal that in our study Bing visual search gives quite better and more relevant responses in context to an event as compared to google image search responses. We will discuss the detailed comparative study of both image search engines in the later section.

In this study, we have considered 'n' events and there are 'm' multimedia posts concerning each event. For each multimedia post, there are 'r' users showing their point of expression by

[22] See it, search it | Bing Visual Search

[23] https://images.google.com/

posing ‘k’ claims. We can show the complete scenario and relationship between different object modules of our system.

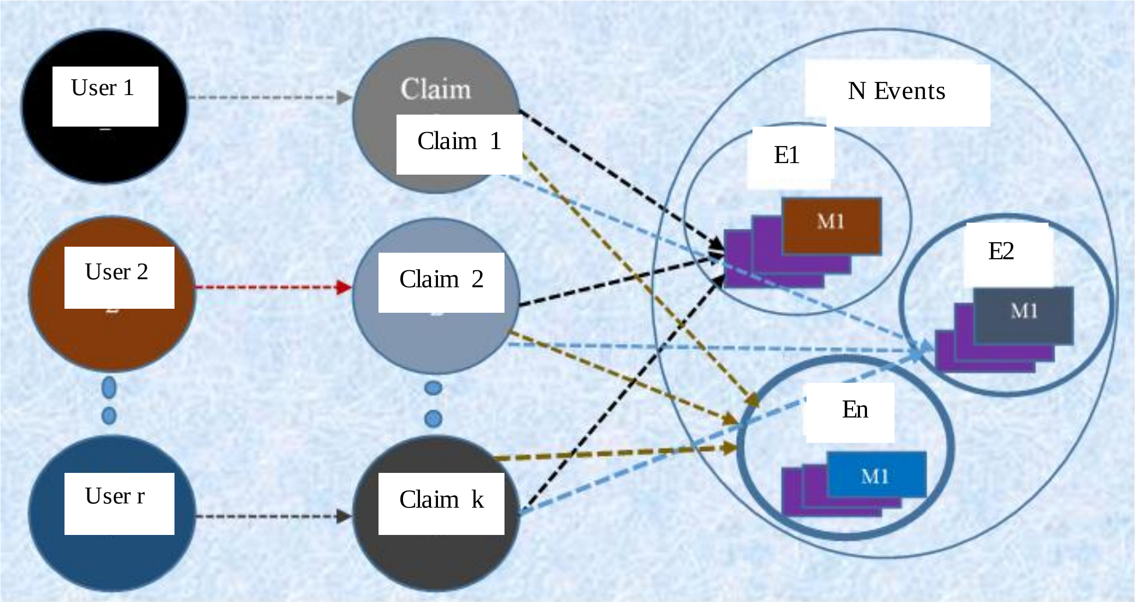


Fig.4.1.1. The figure represents the relationship between user, claims, events and an image.

The graphical representation of our system which is a group of users, claims, events, and an image is shown in Fig.4.1.1. The graph clearly shows the relationship among them, where there are a set of *“r”* users posting different opinions about a specific multimedia post related to some event. There are “*n*” events, each event accompanying a *“k”* multimedia post. Opinions give a set of claims that a user is thinking about the specific event and expressing their thought to represent a given situation. Most of the time on social media people share thoughts without verification, just that post goes viral, people are supporting the given news. While posting any multimedia post, there can be multiple possible causes that can be applied with respect to a human point of expression.

- The user is in support of the claim and confident with the accompanying multimedia item(image) from an event. This we termed as confident claims $CON^{(i)}$.
- The user is in doubt with the claim and with the accompanying multimedia item(image) from an event. This we termed as doubtful claims $DB^{(i)}$.
- The user is not in support of the claim and not confident with the accompanying multimedia item(image) from an event. This we termed as unsupportive claims $UNS^{(i)}$

By understanding the human point of expression, we can evaluate the uncertainty score of the claims provided by users on a specific event and can observe user’s expressions using Eq.4.1.1. We can evaluate the uncertainty score. The uncertainty score can be calculated as the Boolean sum of DB and UNS value for the ith claim. There is a list of phrases and a corpus of words is created from the empirical analysis of collected data. For the doubtful claims, we are analyzing whether the tweet contains any question marks. Question marks are an effective way of identifying the user expression that he/she is in doubt with the given accompanying multimedia content and it represents the uncertainty in their opinion. If any question mark has been identified in the tweet, the DB value will be 1 and 0 otherwise.

$$uncertanity\_score(CS) = (DB^{(i)} + UNS^{(i)}) \tag{4.1.1}$$

4.1.2 Evidential Clues for the Verification of Misleading Multimedia Content

Selecting and incorporating the right set of features and input parameters plays an important role in the better performance of the model. The effective features have been extracted from the multimedia post that leads to give an efficient clue for the prediction of misleading content. The evidential clues have been extracted from two parts 1) Tweet part and 2) Tweet+ Image part. These two parts have been discussed in detail.

- **Evidence Collection from Tweet Part:**

In this section, we are going to discuss the factor that is considered for the collection of clues from the posted tweet/claim. In this study, we have considered multimedia posts with a claim/tweet and the accompanying multimedia item(image). The available tweets are in multilingual form, to understand the semantics, language translation has been applied using google trans library of python. Google trans is a free and unlimited python library that implemented Google Translate API[24]. After analyzing the tweet, it has been observed that a pattern of question marks (Trace of doubt) and a trace of false phrases (Trace of fake) can be an efficient clue for the prediction of false information described as follows. Trace of doubt is one of the patterns that widely identified in the human expressing pattern when he/she is in doubt regarding what they are posting and not sure regarding the post. After analyzing the dataset, we built a corpus having phrases concerning to trace of doubt. We have observed that the prominently used words for expressing doubts are $\{is\ it,\ is\ that,\ Not\ sure, ?\}$. The return value is binary, if it returns 1 means that the tweet expressing doubt, otherwise 0. Here we have represented the trace of doubt with the term DB which has been discussed in the later section. Whereas, Trace of fake is another pattern we have analyzed in the tweets, where the user itself shows the expression of fake, and showing that they are not in support with the claim. We have

[24] https://pypi.org/project/googletrans/

built a corpus*{'Malware', 'Beware', 'scam', 'fishy', 'phishing', 'funny', 'Not', 'ambiguous',' false', 'misleading', 'inaccurate', 'rumor', 'rumour', 'fool', 'fooled', 'not correct', 'wrongly', 'wrong',' misidentified', 'fake news', 'falsely', 'incorrect', 'memes', 'catchy', 'bogus', 'fabricated', 'forged', 'fraudulent', 'artificial', 'erroneous', 'faulty', 'improper', 'invalid', 'invalid', 'mistaken', 'unreal', 'untruthful', 'fishy', 'illusive', 'imaginary', 'lying', 'misrepresentative', 'falsity', 'falsification', 'fabrication', 'falsehood', 'hoax', 'incorrect', 'not real', 'not true', 'fishy', 'illusive', 'imaginary', 'lying', 'misrepresentative', 'falsity', 'misreport', 'deception', 'falsification', 'lie', 'scandal', 'misinformation', 'misleading', 'not dead', 'death rumor', 'not known', 'no proof', 'no scientific evidence', 'denied', 'deny', 'unverified', 'myth'*} of prominently used words pattern in the tweets for representing the trace of fake. Here we have represented the trace of fake with the term UNS. If any of the word patterns have been detected in the tweet it will return 1 otherwise 0.

- **Evidence Collection from both Tweet and Image Part**

In this section we are going to cover the set of evidence or clues that have been collected from the tweet as well as from an image, before going to discuss the clues related to an image, let's first discuss how we can process an image to retrieve relevant knowledge? In our proposed idea, any multimedia post attached with an image is processed as follows, the associated image is given as an input to the image search engines (i.e., Google Image search and Bing visual search) and each search engine returned relevant instances matching with the image. So, in this case, the verification of result responses, whether they are related to the search query is not necessary, because here by default we are getting only those instances, having correlated images. The retrieved titles from each image further are used to gather clues. The following measures that are utilizing both tweets and images for gathering efficient clues for the prediction of misleading information.

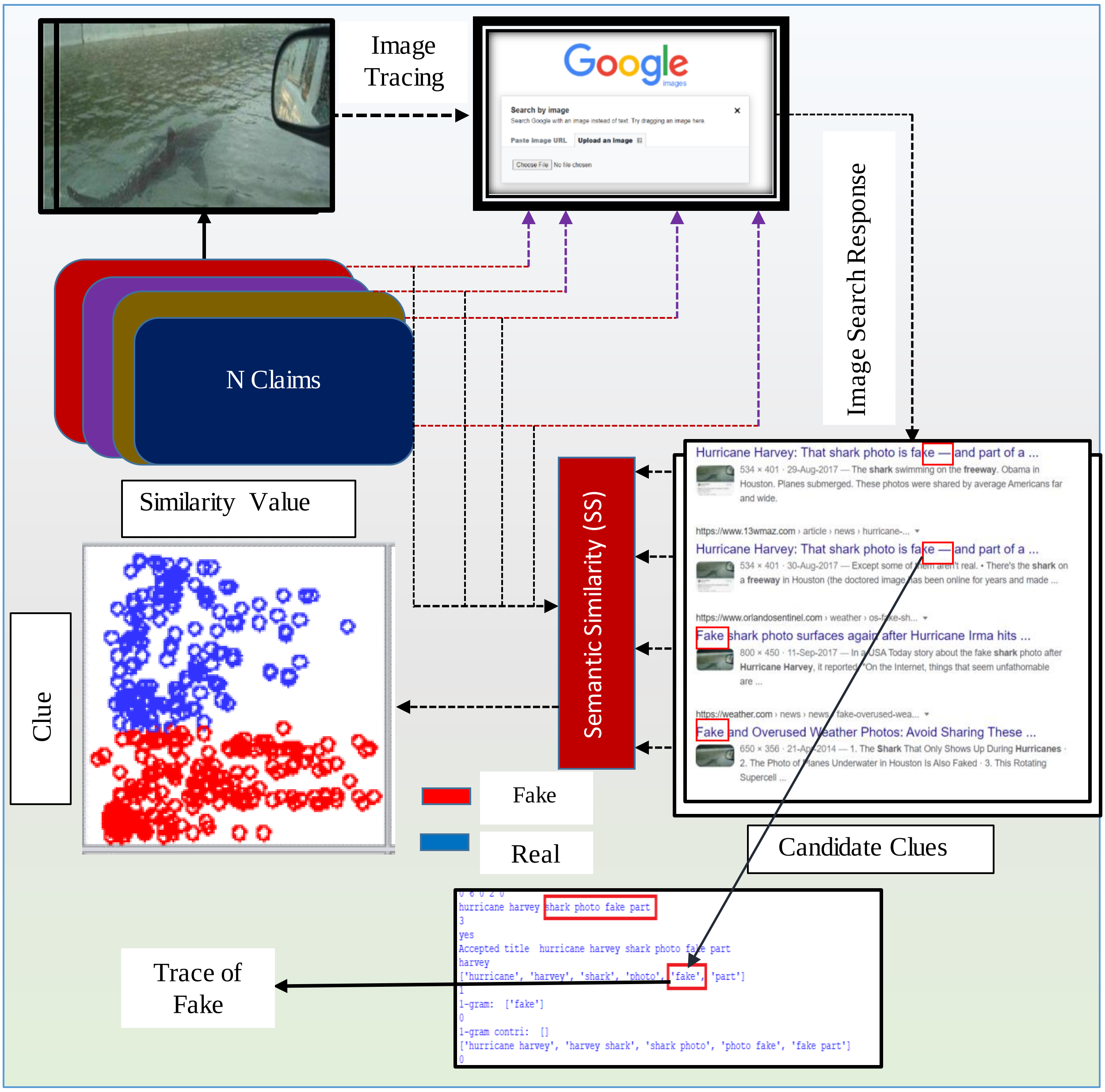
The first one is the semantic similarity measure, the semantic similarity between a tweet and a title retrieved from an image search response ranges from 1 to 10(Top 10 titles) has been calculated. The semantic-text-similarity library of python is an easy-to-use interface to fine-tuned BERT models for computing semantic similarity[25]. This semantic similarity can be one of the good measure to compute how similar the two sentences contextually. This will also reveal whether the posted claim/tweet faithfully represents the accompanying image or not. The semantic BERT similarity maps batches of sentence pairs to the real-valued scores in the

[25] semantic-text-similarity · PyPI

range [0,5]. From the empirical analysis of the similarity value in the dataset, we decide the threshold values that reflect whether the tweet and title are represented in the same context or contradictory/not matched.

Table 4.1.1, shows the set of possible cases that can be applicable and by empirical analysis on Bert-semantic similarity score we have decided the threshold value T, if $T < 1.3$ it has been observed that the given claim and title point of expression are not in the same context and contradictory or not matched to each other. For example, suppose the query is “This image is NOT MH370, this is an image from the incident of a plane crashed in Sicily on 6Ogos2005 #PrayForMH370” and the retrieved title is “Atr72 air disaster, Bari remembers 16 victims”. The computed semantic similarity value is 1.03 which is less than the threshold value T, and it clearly represents that the title and the query are represented in a different context and whereas, if the T>= 1.3 it shows that the query and tweet are represented in the same context, for example, the query is “*This image is NOT MH370, this is an image from the incident of a plane crashed in Sicily on 6Ogos2005 #PrayForMH371*” and the title is “*Serious! - Pictures of MH370 Crashed at Sea This Is Fake UPDATES*” have T value 2.125, which is more than 1.3. In addition to this, the trace of fake has also been checked concerning each query and title that whether they are reporting some expression of fake. Three cases can be possible here as shown in Table 4.1.1.

The first case is when the Query itself reporting news as fake, while clue is not reporting the trace of fake and in contradiction or not matched, while the second case says that the Query is not reporting the trace of fake, while clues are reporting and in contradiction. Whereas, the third case is Query and clue both are reporting the news as fake and in support of each other. In Fig.4.1.2, the Process describes how semantic similarity value between query and clue can be an effective factor in classifying fake and real. This set of features are passed to the machine learning model for the prediction of misleading posts as shown in Fig.4.1.3.



Hurricane Harvey: That shark photo is fake — and part of a ...
534 × 401 · 29-Aug-2017 — The **shark** swimming on the **freeway**. Obama in Houston. Planes submerged. These photos were shared by average Americans far and wide.

https://www.13wmaz.com › article › news › hurricane-...
Hurricane Harvey: That shark photo is fake — and part of a ...
534 × 401 · 30-Aug-2017 — Except some of them aren't real. • There's the **shark** on a **freeway** in Houston (the doctored image has been online for years and made ...

https://www.orlandosentinel.com › weather › os-fake-sh...
Fake shark photo surfaces again after Hurricane Irma hits ...
800 × 450 · 11-Sep-2017 — In a USA Today story about the fake **shark** photo after **Hurricane Harvey**, it reported, "On the Internet, things that seem unfathomable are ...

https://weather.com › news › news › fake-overused-wea...
Fake and Overused Weather Photos: Avoid Sharing These ...
650 × 356 · 21-Apr-2014 — 1. The **Shark** That Only Shows Up During **Hurricanes** · 2. The Photo of Planes Underwater in Houston Is Also Faked · 3. This Rotating Supercell ...

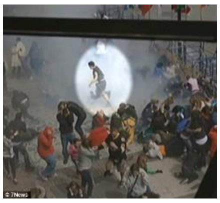
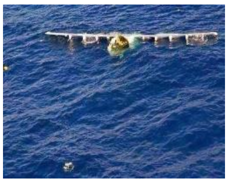
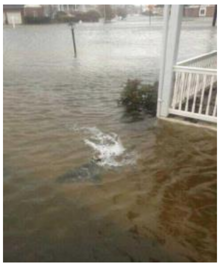
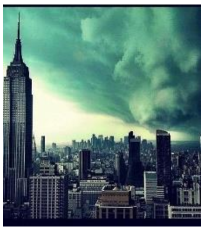
```
0 6 0 2 0
hurricane harvey shark photo fake part
3
yes
Accepted title  hurricane harvey shark photo fake part
harvey
['hurricane', 'harvey', 'shark', 'photo', 'fake', 'part']
1
1-gram:  ['fake']
0
1-gram contri:  []
['hurricane harvey', 'harvey shark', 'shark photo', 'photo fake', 'fake part']
0
```

Fig.4.1.2. Process describing how semantic similarity value between query and clue can be an effective factor classifying fake and real.

Table 4.1.1. The set of possible fake cases that can be applicable.

BERT Semantic Similarity value(T)	Identified Fake Cases	Query False Phrases	Clue False Phrases
T<1.3	Context is not same	-	-
T>=1.3	Query/ Root itself reporting a news as fake, while clue is not reporting the trace of fake and in contradiction	Yes	No
T>=1.3	Query/ Root is not reporting the trace of fake, while clues are reporting and in contradiction.	No	Yes
T>=1.3	Query/ Root and clue both are reporting the news as fake and in support of each other.	Yes	Yes

Table 4.1.2. Image search result responses from Microsoft Bing and google image search.

Images	Microsoft Bings Image Search Responses	Google Image Search Responses
	• Now: FBI Hunts Suspect in Boston | Bomb Attack. • Common Cents: Are these photos of the Boston Bombing Suspect? • Photo of Boston Bomber Caught on camera | TODAY'S JOBS |. • Boston Marathon suspects Archives • Who of these men is the Boston Marathon Attacker**?** Possible Suspect in.	• event web apis mdn • event meaning Cambridge English dictionary • search results signals az. • digital vigilantism boson marathon bombing • boston marathon bombings latest arrest made.
	• Atr72 air disaster, Bari remembers the 16 victims • Cape Gallo air disaster, 11 years ago the Atr72 tragedy.	• Crash pilot who paused to pray is convicted | Reuters
	• Is that picture real or **fake?** - Is that right? • 20 Epic **Fake** Pictures that Have **Fooled** the Whole World | Shark swimming. • The Big Apple has lots of sharks. But real ones in the neighbourhood. • Super Storm Sandy Sharks swimming down New Jersey street. • Hurricane Irene: 'Photo' of shark swimming in street is **fake** | Shark .	• 54 Super storm sandy ideas | sandy, storm, hurricane sandy. • 72 Crazy shit ideas | hurricane sandy, natural disasters, photo. • 7 Sandy ideas | sandy, hurricane sandy, hurricane pictures. • These Viral Shark Photos from Hurricane Matthew Are, Once. • **Fake** and Overused Weather Photos: Avoid Sharing These.
	• **Is that** really a picture of Hurricane Sandy descending on New York**.?** • NY City | Hurricane pictures, New York photos, New york | • Hurricane Sandy 2012: 10 Amazing Photos of the Storm's Path Through New. • These Are **NOT** Photos From Hurricane Sandy (No Matter What The Internet.	• Internet Awash in #**Fake** Sandy Photos. Have You Shared Any? • 22 Viral Pictures That Were Actually **Fake** | Hurricane pictures ... • Example of **fake** picture of stormy New York skyline used in ...

Whereas, the second way is to detect complex patterns from tweet and image search responses using Bi-directional LSTM. To Extract the complex hidden representation from tweets and image search responses, a Bi-directional Long short-term memory network (Bi-LSTM) a special type of RNN competent in learning long dependencies is utilized in our proposed work as shown in Fig.4.1.3. An RNN has an internal state whose output at every time step can be expressed in terms of the previous time step. However, RNNs suffer from the problem of vanishing and exploding gradients[26] and this leads to the model learning inefficient dependencies between words that are a few steps apart. To overcome this issue, the LSTM extends the basic RNN by storing information over long periods using its memory units and efficient gating mechanisms. LSTM is a special type of RNN competent in learning long-term dependencies and they are providing an efficient solution to address the vanishing gradient problem. In LSTM-RNN the hidden layer of basic RNN is replaced by an LSTM cell. LSTM is prominent as they utilize various different gates in their architecture that help in learning how and when to forget and when not to. Another variant of RNN is Bi-directional LSTM, where you feed the learning algorithm with the given data in two ways once from beginning to the end and once from end to beginning. From the study, it has been observed that for a large text sequence prediction and text classification, Bi-directional LSTM was found to be an effective and evident approach, which takes a step through the input sequence in both directions at the same time.

The proposed misleading content detection model is based on Bi-directional LSTM – recurrent neural network. The claim and the image search responses(Titles) corresponding to each image are first pre-processed (removing stop-words, stemming, lemmatization, removing URLs, punctuation). Concerning each image, there are n responses retrieved (n titles). A binary label is set to each title as 1 for fake news and 0 for real news corresponding to the individual query. The titles retrieved from image search responses and the corresponding query are turned into a space-separated padded sequence of words. These sequences are further split into tokens. One hot vector encoding embeddings is utilized to represent each word by the real value number. The embeddings are then passed to Bi-directional LSTM Model to detect complex hidden patterns/features from the text. The transformed vector represented data is partitioned into train, validation, and test data. The training is carried out on the build corpus of queries and titles concatenated with a space. Validation data set is used for fine-tuning the model. Further, the

[26] Recurrent Neural Networks (RNN) - The Vanishing Gradient Problem - Blogs SuperDataScience - Machine Learning | AI | Data Science Career | Analytics | Success

test data is used to know the predicated label of news content (query + title) based on trained model.

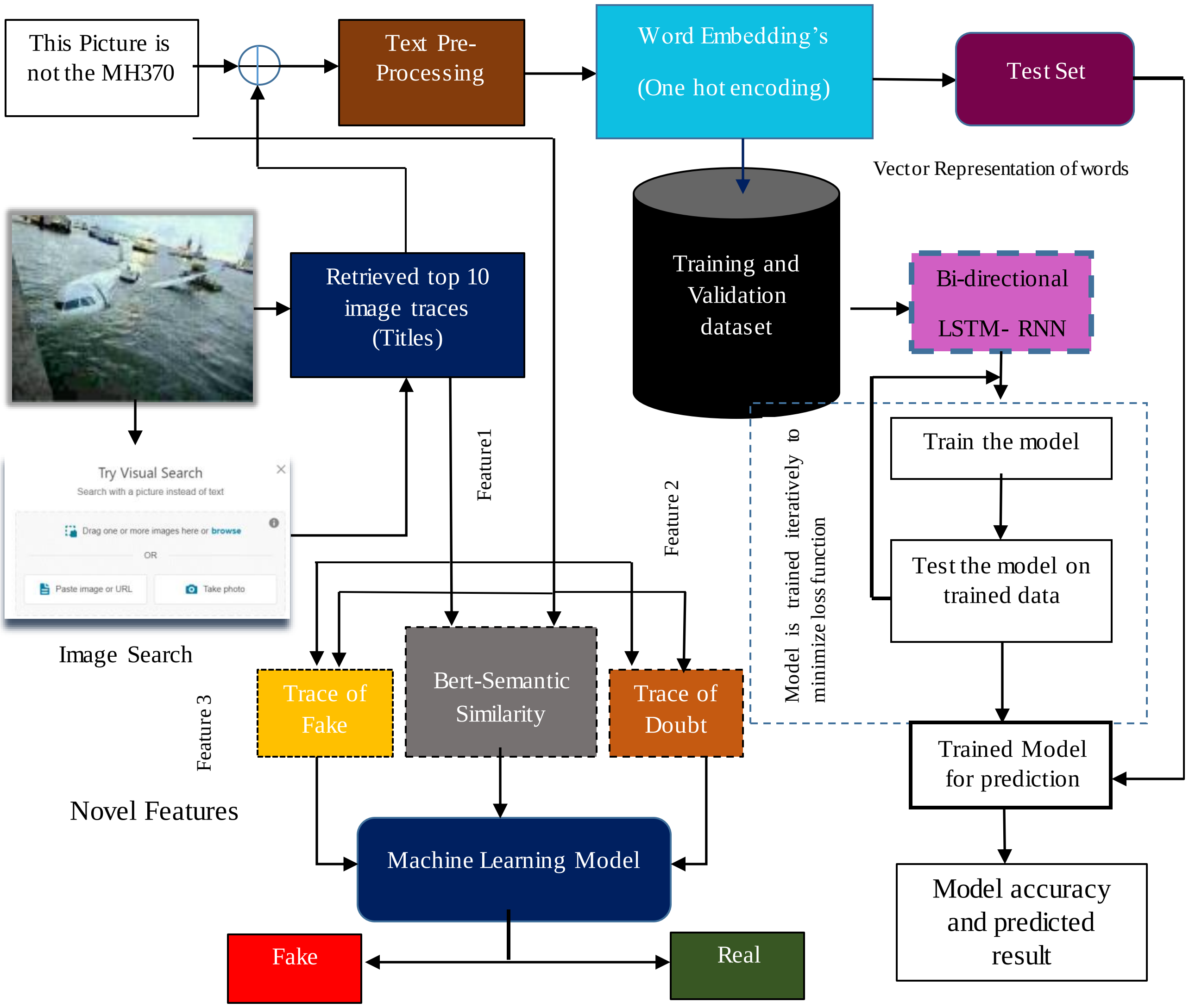

This Picture is not the MH370
Text Pre-Processing
Word Embedding's
(One hot encoding)
Test Set
Vector Representation of words
Training and Validation dataset
Bi-directional
LSTM- RNN
Retrieved top 10 image traces (Titles)
Try Visual Search
Search with a picture instead of text
Drag one or more images here or browse
OR
Paste image or URL
Take photo
Image Search
Feature1
Feature 2
Feature 3
Model is trained iteratively to minimize loss function
Train the model
Test the model on trained data
Trained Model for prediction
Model accuracy and predicted result
Trace of Fake
Bert-Semantic Similarity
Trace of Doubt
Novel Features
Machine Learning Model
Fake
Real

Fig.4.1.3. The Proposed architecture to Detect complex patterns from tweet and Image search responses using Bi-directional LSTM (Deep learning) and machine learning models.

To minimize the loss function, the model is trained iteratively to improve accuracy. The binary cross-entropy loss is considered to detect misleading multimedia posts in the proposed model. The Adam optimization algorithm is used to improve the performance of the model.

4.1.3 Experiments and Results

In this section we are going to discuss experimental analysis and later demonstrates the results we have achieved by applying our proposed approach for the detection of misleading content on social media. We then briefly discuss some state-of-the-art techniques used in this field and lastly show the comparative analysis with baselines to validate the performance of our model.

- **Dataset**

In this section, we discuss the dataset that has been employed to evaluate the performance of the model. One of the prominently used standard datasets is the Mediaeval verifying multimedia Use challenge[27]. The task was aimed to predict the misleading multimedia content on social media. The dataset is comprised of a set of Twitter posts having tweets associated with multimedia items. The VMU(Verifying Multimedia Use 2015) is a publicly available dataset [89] on GitHub[28] . The dataset incorporated social media posts having ~400 images (176 cases of real and 185 cases of misleading images) associated with 5,008 real and 7,032 fake tweets concerning 11 events (Boston Marathon bombing, Hurricane Sandy, etc.). Whereas, for the evaluation of the model test dataset has been used, comprises of total 50 images (17 real and 33 misleadings) and is associated with 1,217 real and 2,564 fake tweets. Table 4.1.3 shows the detailed description of the VMU 2015 dataset. As in this study, our main focus is on the images and textual information, that's why the tweets that are associated with videos are filtered out.

Table 4.1.3. The table represents the detailed description of the VMU 2015 Dataset.

Event Name	Real Images	Real Tweets	Fake Images	Fake Tweets
Hurricane Sandy	148	4,664	62	5,559
Boston Marathon Bombing	28	344	35	189
Sochi Olympics	-	-	26	274
MA flight 370	-	-	29	501
Bring Back Our Girls	-	-	7	131
Columbian Chemicals	-	-	15	185
Passport Hoax	-	-	2	44
Rock Elephant	-	-	1	13
Underwater Bedroom	-	-	3	113
Livr mobile app	-	-	4	9
Pig fish	-	-	1	14
Total	176	5,008	185	7,032

The study is conducted on the machine as well as deep learning approaches by utilizing tweets, images, and the combination of Tweet and images. In the following subsection, we separately

[27] Verification (New!) (multimediaeval.org)
[28] https://github.com/MKLab-ITI/image-verification-corpus.

discuss the effectiveness of employing tweet only, image only, and both (image and tweet) the ways as well as analyze the performances with respect to each case.

- **Performance Evaluation on Machine learning models**

The effectiveness of the proposed method has been evaluated by assessing the novel features employed for the prediction of misleading content. The five-set of novel features as discussed in the earlier section (Trace of fake concerning to query, Trace of fake concerning to titles, Trace of doubt concerning to query, Trace of doubt concerning to titles, the semantic similarity between title and a query) with respect to tweet and images are fed into machine learning model to validate how significant these features in improving the performance of the model. The titles concerning to an image are retrieved using Microsoft Bing visual search and the performance of the model has been evaluated using TP rate, FP rate, Precision, Recall, F1 score, and accuracy as shown in Table 4.1.4. From Table 4.1.4, it can be observed that Random forest and Linear SVM performing better and outperforming all other classifiers with an F1 score of 0.978.

Table 4.1.4. Effectiveness of the proposed model using machine learning methods

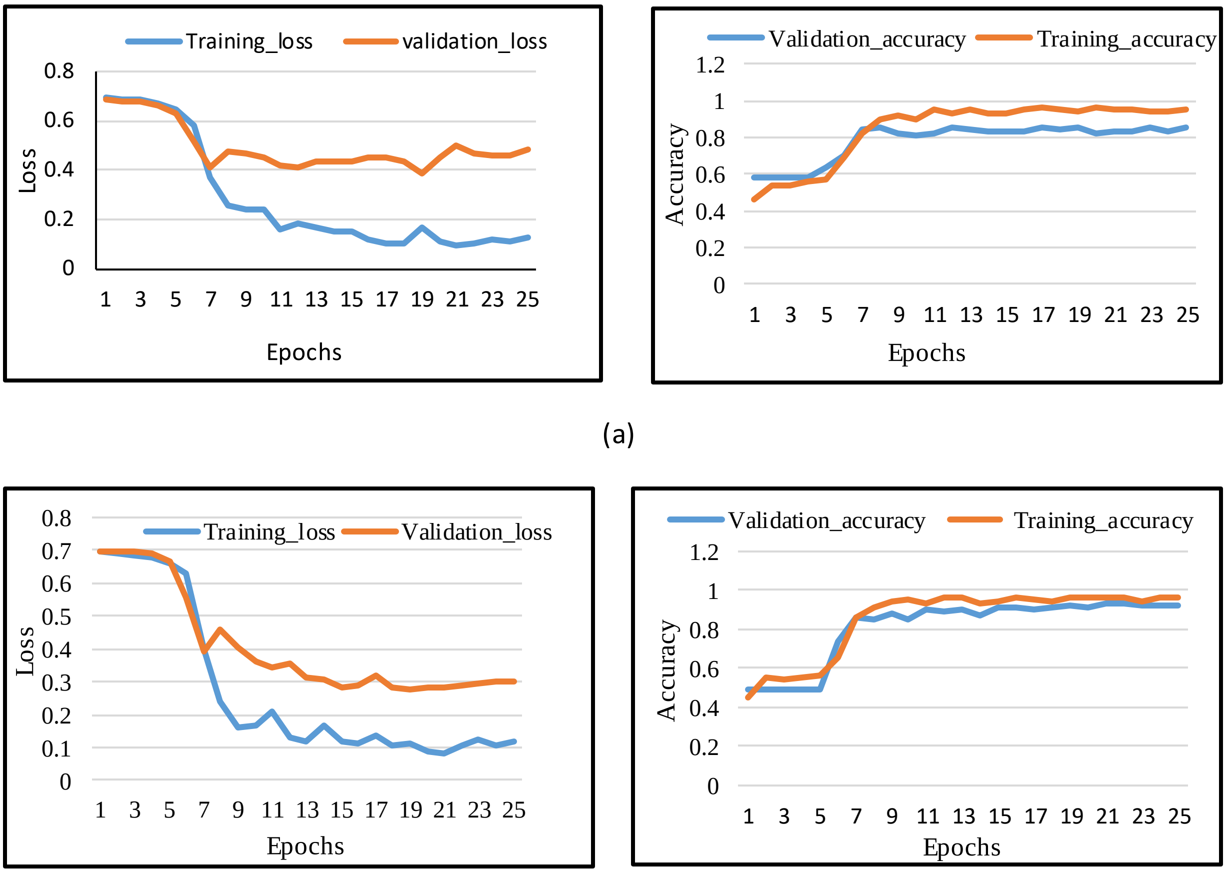
Classifier	Performance Measures					
	TP Rate	FP Rate	Precision	Recall	F- Measure	Accuracy
Random Forest	0.978	0.019	0.979	0.978	**0.978**	97.81
Logistic Regression	0.970	0.026	0.971	0.970	0.970	96.99
Naïve Bayes	0.929	0.062	0.936	0.929	0.929	92.89
Linear SVM	0.978	0.026	0.979	0.978	**0.978**	97.81
K-Nearest Neighbour	0.967	0.029	0.968	0.967	0.967	96.72

- **Performance Evaluation on Deep learning models**

To extract complex hidden representation/ features from textual data, the Bi-directional LSTM model has been employed as discussed in later sections. Here, the performance of the proposed model has been evaluated by employing two prominently used image search engines "Google search and Microsoft Bing visual search" for the retrieval of image search responses. It has been observed that getting effective search responses concerning an image is one of the crucial measures in improving the performance of the model. The performance of the model degrades if significant responses/titles have not been retrieved. To validate this point, the comparative study has been performed by employing two prominent image search engines for the retrieval of image search responses on the Mediaeval dataset, the provided study is when only image search responses(titles) are passed to the Bi-directional LSTM model. One of the examples is represented with respect to an event *"Boston Marathon Bombing"* as shown in Fig.4.1.4. The loss and accuracy curve corresponding to the number of epochs is shown to demonstrate the performance of the model. It has been observed from Fig.4.1.4(a), that we are achieving the validation accuracy of 0.93 when utilizing Microsoft Bing as an image search engine which is quite good and better in compare when utilizing google chrome image search results (validation accuracy of 0.85) on *"Boston Marathon Bombing*" when reaching 25th epoch as shown in Fig.4.1.4(b). From the complete observation, we found that utilizing Microsoft Bing image search is better to improve the performance of our model on our data, that's why incorporated the same for the further analysis. The other set of experiments has been performed on the overall dataset, the provided study is when only image search responses(titles) are passed to the embedding layer and then further passes to the Bi-directional LSTM model. It can be seen from Fig.4.1.5, that we are able to achieve a validation accuracy of 0.86, and loss is reduced to 0.41. In order to improve the performance of the model, instead of just passing image-based clues, the claim is also incorporated to get effective features. The Tweet and Images search responses are concatenated separately with space and passes to the model. It has been observed from Fig.4.1.6. that there is a significant improvement we achieved in this case, we got a validation accuracy of 0.99 and loss is almost reached to 0.

4.1.4 Comparative study with state-of-the-art approaches

The comparative study has been performed with the other state-of-the-art methods to evaluate the performance of our proposed approach. We compare the techniques applied on Mediaeval VMU dataset 2015 as discussed in earlier sections. From Table 4.1.5, it can be observed that the proposed method outperforms the state-of-the-art technique on the same dataset. The main performance measure that has been used for the comparison is F1-score and approaches are compared against their best run. Among all other methods (these include the method proposed by [90], [91], [92], and [93], our method outperforms with an F1 score of 0.99 using the Bi-directional LSTM model and give the best run when considering both tweet and image. However, it gives an F1-Score of 0.86 when utilizing only image-based evidence. The authors of [79], employed supervised machine learning methods for evaluating the performance of their model, where they achieved an F1-score of 0.932 and 0.935 with Logistic Regression and Random Forest respectively. Whereas, by employing our proposed novel features, we achieved an F1-Score of 0.978 and 0.970 with random forest and logistic regression respectively.



(a)

(b)

Fig.4.1.4 The training and validation loss as well as accuracy curve corresponding to no. of Epochs for Boston Marathon Bombing. (a) Google Chrome (b) Microsoft Bings (Image only).

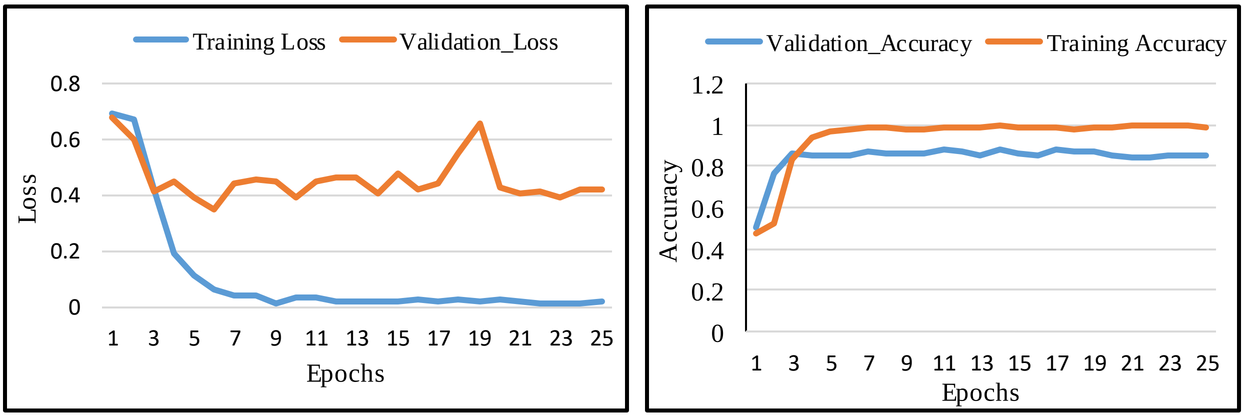


Fig.4.1.5 The training and validation loss as well as accuracy curve corresponding to no. of Epochs for overall dataset (VMU 2015) using Microsoft Bings (Image only).

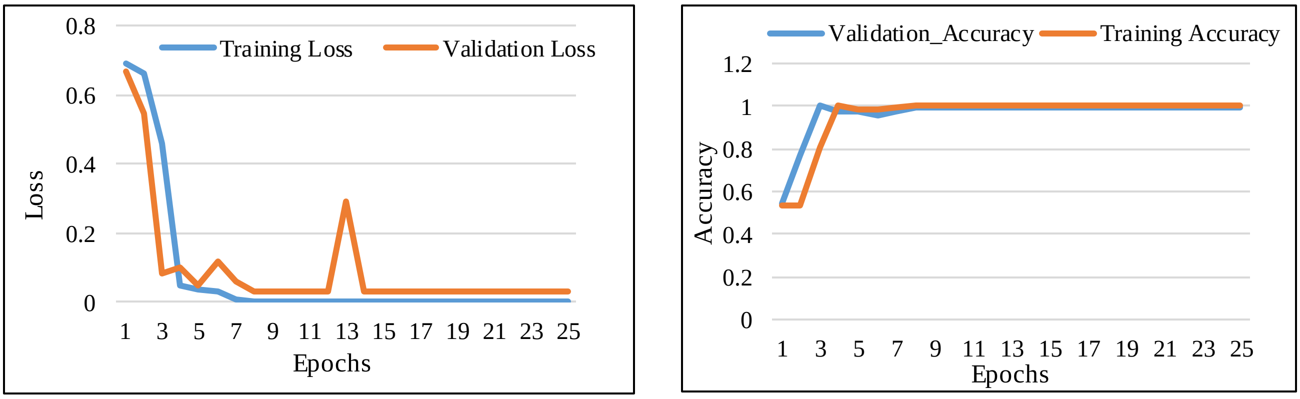


Fig 4.1.6. The training and validation loss as well as accuracy curve corresponding to no. of Epochs for overall dataset (VMU 2015) using Microsoft Bings (Image + Tweet only).

Table 4.1.5. The Comparative study between the proposed method and the state-of-the-art method on the mediaeval VMU 2015 dataset.

Ref	Method	Type of Input	Performance Measure			
			Precision	**Recall**	**F1-Score**	**Accuracy**
[90]	UoS-ITI	Tweet+ Image	-	-	0.830	-
[91]	MCG-ICT	Tweet+ Image	-	-	0.942	-
[92]	CERTH-UNITN	Tweet+ Image	-	-	0.911	-
[79]	Logistic Regression	Tweet+ Image	-	-	0.932	-
[79]	Random Forest	Tweet+ Image	-	-	0.935	-
Our Method	LSTM	Image only	0.86	0.86	0.86	0.86
	LSTM	Tweet+ Image	0.99	0.99	**0.99**	0.99
	Random Forest	Tweet + Image	0.979	0.978	**0.978**	97.81
	Logistic Regression	Tweet+ Image	0.971	0.970	**0.970**	96.99
	Naive Bayes	Tweet+ Image	0.936	0.929	0.929	92.89
	Linear SVM	Tweet+ Image	0.979	0.978	**0.978**	97.81
	K-Nearest Neighbour	Tweet + Image	0.968	0.967	0.967	96.72

4.2 Detecting fraudulent content accompanying Video

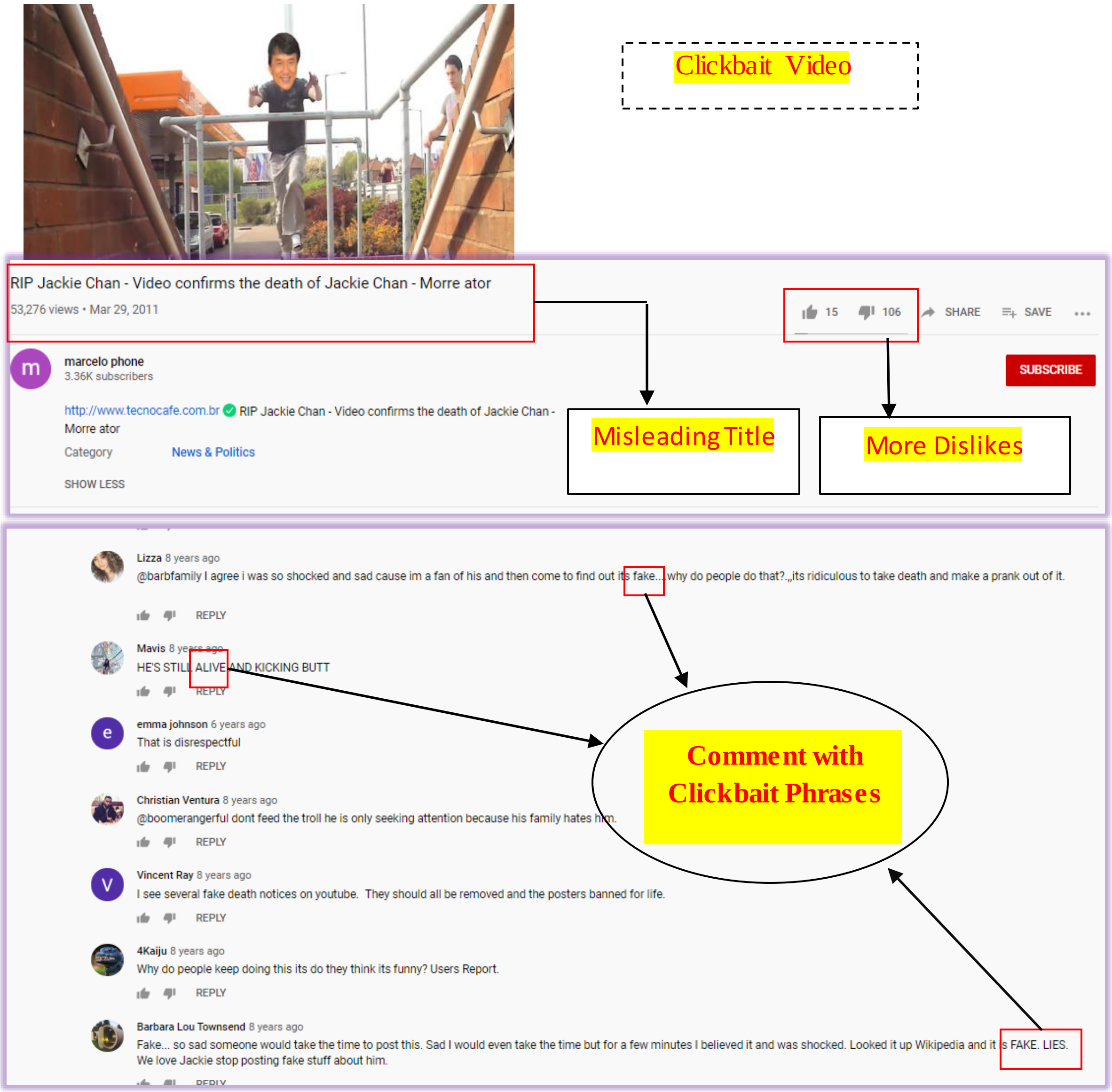
In the previous section, we have discussed our proposed technique of detecting fraudulent content accompanying an image. Like images, videos are also prominently used with the content to express a situation. In this section, we discussed our proposed method of detecting

FC accompanying video. There are many different forms of fraudulent content as we discussed in the earlier sections. In this section, we addressed the problem concerning clickbaits. Clickbait is one of the forms of false content, purposely designed to attract the user's attention and make them curious to follow the link and read, view, listen to the attached content. The teaser's aim behind this is to exploit the curiosity gap by giving information within the short statement. Still, the given statement is not sufficient enough to satisfy the curiosity without clicking through the linked content and lure the user to get into the respective page via playing with human psychology and degrades the user experience. To counter this problem, we develop a Clickbait Video Detector (CVD) scheme. The scheme leverages to learn three sets of latent features based on User Profiling, Video-Content, and Human Consensus, these are further used to retrieve cognitive evidence for the detection of clickbait videos on YouTube. The first step is to extract audio from the videos, which is further transformed to textual data, and later on, it is utilized for the extraction of video content-based features. Secondly, the comments are analyzed, and features are extracted based on human responses to the posted content. Lastly, user profile-based features are extracted. Finally, all these features are fed into the classifier. The proposed method is tested on the publicly available fake video corpus [FVC], [FVC-2018] dataset, and a self-generated misleading video dataset [MVD]. The achieved result is compared with other state-of-the-art methods and demonstrates superior performance. Clickbait’s are purposely designed to attract the user's attention and make them curious to follow the link and read, view, listen to the attached content. In 1994, George Loewenstein has explained clickbait, *“as the information gap theory of curiosity*” [94]. We followed this definition and define clickbait “*as the information gap theory of curiosity, that play with human psychology, to lure the user to view a content that does not faithfully represent the claim it presenting and degrades the user experience*”. Whereas, “*Non-click baits can be defined as the content that is presenting the real news and faithfully giving the same picture of content to the viewer, that it claiming for*”. The work provides a detailed description of clickbait video detection mechanisms in online social media platforms. Detecting clickbait videos is an intelligent task, as it analyses the video content automatically using clickbait video detection frameworks/tools/plugins, as well as in the future it can also be used as an intelligent warning system that can help to automatically report the credibility of video content to the user. Fig.4.2.1. and Fig.4.2.2. shows the example of clickbait and non-clickbait video in brief. A recent example is of COVID- 19 pandemic, which affects worldwide badly, and there is no shortage of people who are taking this crisis as an opportunity for malicious activities/gaining profit. A lot of health-related misleading information, some of the fake cures are suggested for COVID-19 have been posted by the malicious user, by adding catchy coronavirus headlines to increase their chance of a click, download, or purchase. During this pandemic, people have their eye on any new announcement from the government official or some news that can help to get rid of COVID-19, and the bad guy uses this opportunity to gain more views on their post by adding catchy headlines with the news content that does not faithfully represent the event that it refers to and in this way spreading false information[29]. One of the fake YouTube videos, gone viral with having nearly half a million views falsely said that inhaling hot water from a hairdryer can help to cure the coronavirus[30], which later turned out to be false. The presence of such misleading content over social media makes it more difficult for the user to discriminate credible information from false stories and it leads to making it a challenging area in research, as the spread of clickbait videos not only degrades user experience but also decrease the trustworthiness of video-sharing platforms. Very few works have been reported in detecting clickbait on the video platform. There is a careful analysis required among the features extracted from the video. The current research has not been addressed this problem fully, as they focused only on the content-based solution like the content of the video [95],[96], the image of the thumbnail [97],[98], or text of the title. Most of the text-based clickbait detection methods have been adopted linguistic features[99], or word embedding for the detection of clickbait news headlines, but those solutions cannot be employed to address the clickbait videos, as the only title may not be a reliable indicator, because two videos can share the same title with different content. In the same way, another sort of image-based approach has been employed that focuses on thumbnail features and is not found to be effective in solving the video clickbait detection problem. In this section, we proposed a novel mechanism by introducing three sets of evidential clues, identified and retrieved concerning each video, so that one can easily discriminate unsubstantiated information. The recent work addressed various text-based and comment-based features, however neglecting video speech-based features, as well as user profile features, are also not explored well, which was found to be very effective in detecting clickbait videos. To the best of our knowledge, speech-title similarity has not been explored by the previous research which can be an important clue to solve the problem, when we have two videos with a similar title but different content, the speech is converted into text, and then comparison has been made with the title to check how faithfully the speech is representing the title. Along with it, we have also addressed the problem, when the comments-based features are not retrieved, as the uploader does not allow comments from

[29] https://www.buzzfeednews.com/article/janelytvynenko/coronavirus-fake-news-disinformation-rumors-hoaxes

[30] https://www.bbc.com/news/52124740

the viewers. In that case, relying only on a certain set of features is not effective. To address this case, we have introduced another clue i.e. credible sources, through which we can be able to predict the credibility of the video, the detailed description has been given in the later sections.



RIP Jackie Chan - Video confirms the death of Jackie Chan - Morre ator

53,276 views • Mar 29, 2011

15 106

marcelo phone
3.36K subscribers

http://www.tecnocafe.com.br ✓ RIP Jackie Chan - Video confirms the death of Jackie Chan - Morre ator

Category News & Politics

Lizza 8 years ago
@barbfamily I agree i was so shocked and sad cause im a fan of his and then come to find out its fake....why do people do that?.,,its ridiculous to take death and make a prank out of it.
REPLY

Mavis 8 years ago
HE'S STILL ALIVE AND KICKING BUTT
REPLY

emma johnson 6 years ago
That is disrespectful
REPLY

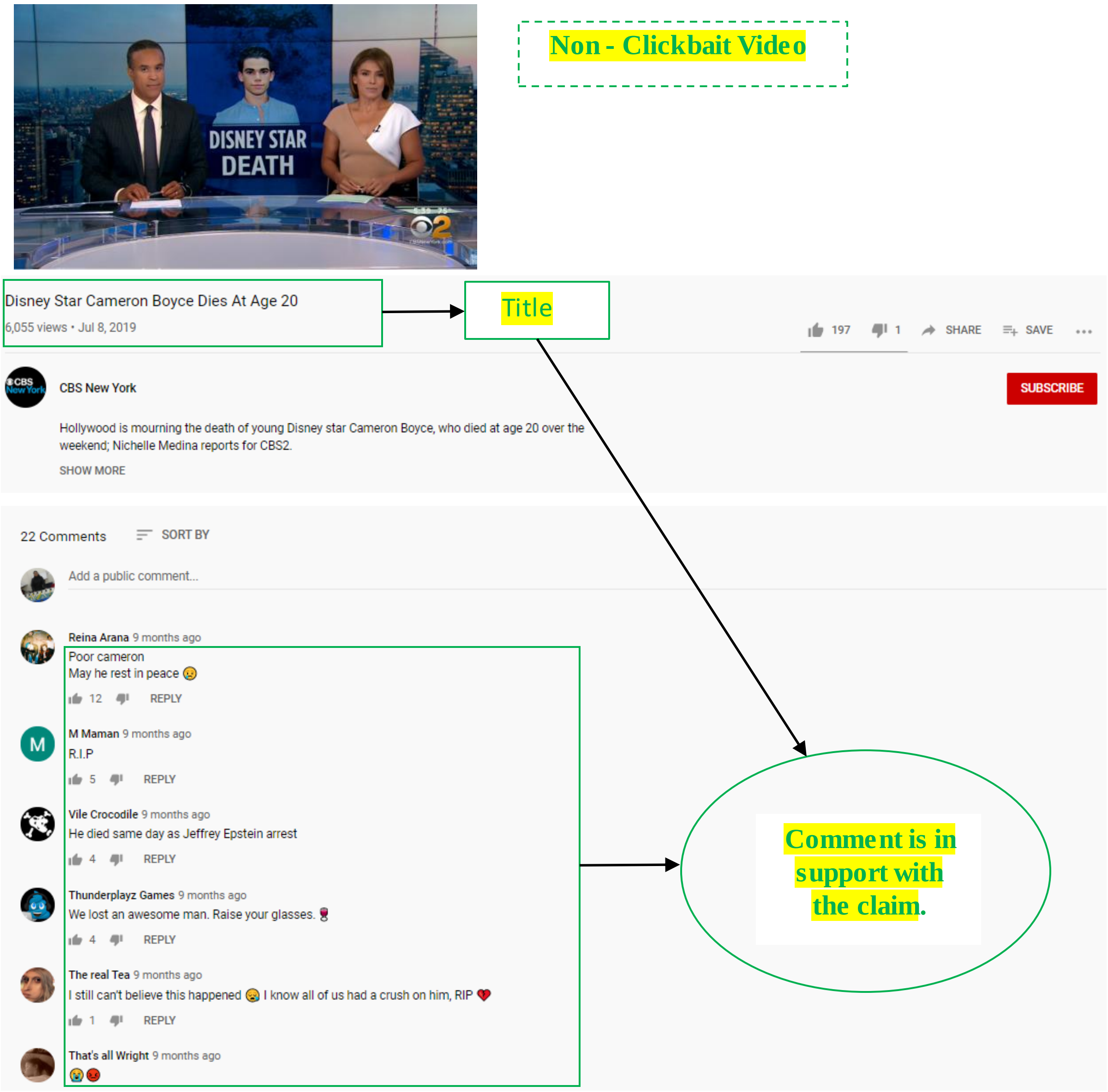
Christian Ventura 8 years ago
@boomerangerful dont feed the troll he is only seeking attention because his family hates him.
REPLY

Vincent Ray 8 years ago
I see several fake death notices on youtube. They should all be removed and the posters banned for life.
REPLY

4Kaiju 8 years ago
Why do people keep doing this its do they think its funny? Users Report.
REPLY

Barbara Lou Townsend 8 years ago
Fake... so sad someone would take the time to post this. Sad I would even take the time but for a few minutes I believed it and was shocked. Looked it up Wikipedia and it is FAKE. LIES.
We love Jackie stop posting fake stuff about him.
REPLY

Fig.4.2.1 Example of Clickbait Video



Disney Star Cameron Boyce Dies At Age 20

6,055 views • Jul 8, 2019

197 1 SHARE SAVE

CBS New York

SUBSCRIBE

Hollywood is mourning the death of young Disney star Cameron Boyce, who died at age 20 over the weekend; Nichelle Medina reports for CBS2.

SHOW MORE

22 Comments SORT BY

Add a public comment...

Reina Arana 9 months ago
Poor cameron
May he rest in peace 😢
12 REPLY

M Maman 9 months ago
R.I.P
5 REPLY

Vile Crocodile 9 months ago
He died same day as Jeffrey Epstein arrest
4 REPLY

Thunderplayz Games 9 months ago
We lost an awesome man. Raise your glasses. 🍷
4 REPLY

The real Tea 9 months ago
I still can't believe this happened 😢 I know all of us had a crush on him, RIP ❤
1 REPLY

That's all Wright 9 months ago
😭😡

Fig.4.2.2 Example of Non-Clickbait

The original contribution of the work can be seen from Fig.4.2.3, where the existing work and the proposed work have been demarcated. The main contribution of the work is as under:

- The major challenge in the area of clickbait detection is a very few public datasets are available for the detection of clickbait videos, and the majority of the datasets are available incorporating clickbait headlines. The self-generated dataset[MVD] has been proposed in this work, which further helps to explore the research in this field.
- Very few methodologies have been proposed that aim to detect clickbait video[100]. This is an emerging field and largely unsolved problem, still very few works have been reported in this area. This gives novelty to our work and also motivates us to provide

an efficient solution for clickbait detection. Three sets of novel features (Video-based, Comment based, and Channel-based) are reported in this work that is found to be efficient and outperform other states-of-the-art on the same dataset.

- To the best of our knowledge, no works have been reported on the concept of tackling videos having commenting off, where the uploader is not allowing users to comment on their video due to which no important clues can be retrieved from the comments section to predict the video as clickbait or real. Top 15 web headlines are fetched and analyzed to get some important evidence about a video and help in efficient prediction.
- It has been observed that many of the existing works have considered video metadata instead of extracting some clues from the video transcripts. To the best of our knowledge, we have first to include the video transcript-based feature to get some informative evidence.
- The comparative analysis has been done with the other existing algorithms. The results clearly show that the proposed model is superior and outperforms the other existing state-of-the-art.

4.2.1 Dataset Creation

One of the significant contributions of this work is dataset creation since very few datasets are available. Hence, a dataset of clickbait's and Non-Clickbait videos have developed by collecting a diverse set of videos using YouTube REST Data API v3. The details include video content (title, likes, dislikes, views, etc.), number of comments, channel details (number of subscribers, registration date, video count, view count). In the field of misleading video detection, very few datasets are available, which leads to giving the aim to build a generalized dataset, incorporating various categories. From the list of 16 most popular videos categories defined by YouTube[31], we have collected 987 videos(474 Clickbait and 513 Non-clickbait). To collect clickbait videos, we manually crawled and annotated each of the 474 videos. We have analyzed some of the channels as well as their following channels that are prominently posted clickbait's, to lure the user to visit their video. Some of the channels[32] that are posted claim make users curious to visit a link for getting more impressions on their video. YouTube has a good check algorithm, for detecting fraudulent videos and also have a blocking mechanism, then also most of the videos are still in their active stage and not removed. From the study, it has been analyzed that most of the clickbait's or hoaxes are posted concerning

[31] https://mediakix.com/blog/most-popular-youtube-videos/
[32] https://www.youtube.com/watch?v=zDa-HzCFolo&t=8s

celebrity's death, which later is found to be false and degrades the user experience. As no correct verification is there provided, this posted news hamper public emotions as well. That's why detecting clickbait video is one of the prominent areas of research.

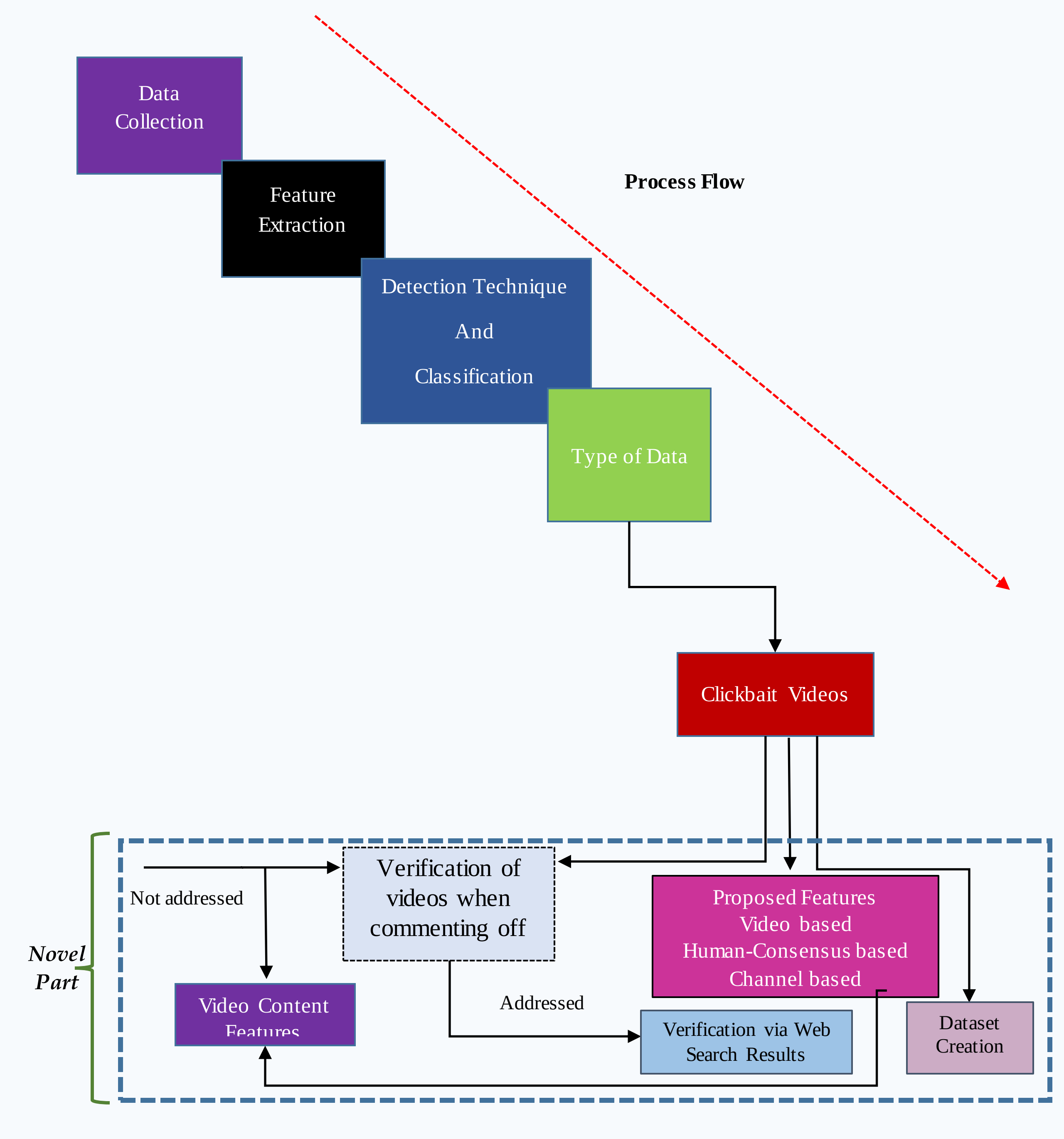

Data Collection
Feature Extraction
Detection Technique And Classification
Type of Data
Process Flow
Clickbait Videos
Novel Part
Not addressed
Verification of videos when commenting off
Proposed Features Video based Human-Consensus based Channel based
Video Content Features
Addressed
Verification via Web Search Results
Dataset Creation

Fig 4.2.3. Flow of Proposed Methodology

To reduce the time of collecting fake videos, the strategy that we follow is viral videos, because they are more likely to generate fake content, with having catchy headlines also prone to make them viral. We manually analyzed the channels/Source of generating these videos[33], additionally the titles are also scrapped and analyzed to get some clickbait phrases like some of them are “Shocking”, “OMG”, “Sad News”, “Bad News”,” Dukhd Khabar”. To direct our search in the correct direction, we try to find out those channels that are following these channels, because it is more likely that the following channel is also posting fake content. Video response also plays an important role while segregating fake/clickbait videos, some of the phrases like “fake video”, “bullshit”, “galat”, “hoax”, “clickbait”, “Alive”, “fake news”,” liar”, “false”, “falsely”,” misinformation”, “rumor”,” clickbait”,” hoax” are used, along with it dislike to like ratio has also been used for further filtering, as it has been observed that clickbait/fake videos received more dislikes compare to likes. For the collection of Non-Clickbait videos, some popular authentic channels have been considered for analysis such as “TEDx Talks”, “Harsh Beniwal”, “Marvel Entertainment” etc. We call this dataset “MVD” (Misleading Video Dataset), and the distribution of the dataset is as given in Fig.4.2.4 and Table 4.2.1. It can be seen from Fig.4.2.4 that 14 different categories have been considered for the dataset creation, whereas in the dataset, the number of clickbait is found more from the “Entertainment” and secondly with “people and blogs” category, as from the manual analysis it has been observed that most of the clickbait’s are prominently available in these categories, while very few videos are considered from the categories(“Pets and Animals”, “Auto and Vehicles”) where the possibility of clickbait generation is quite low.

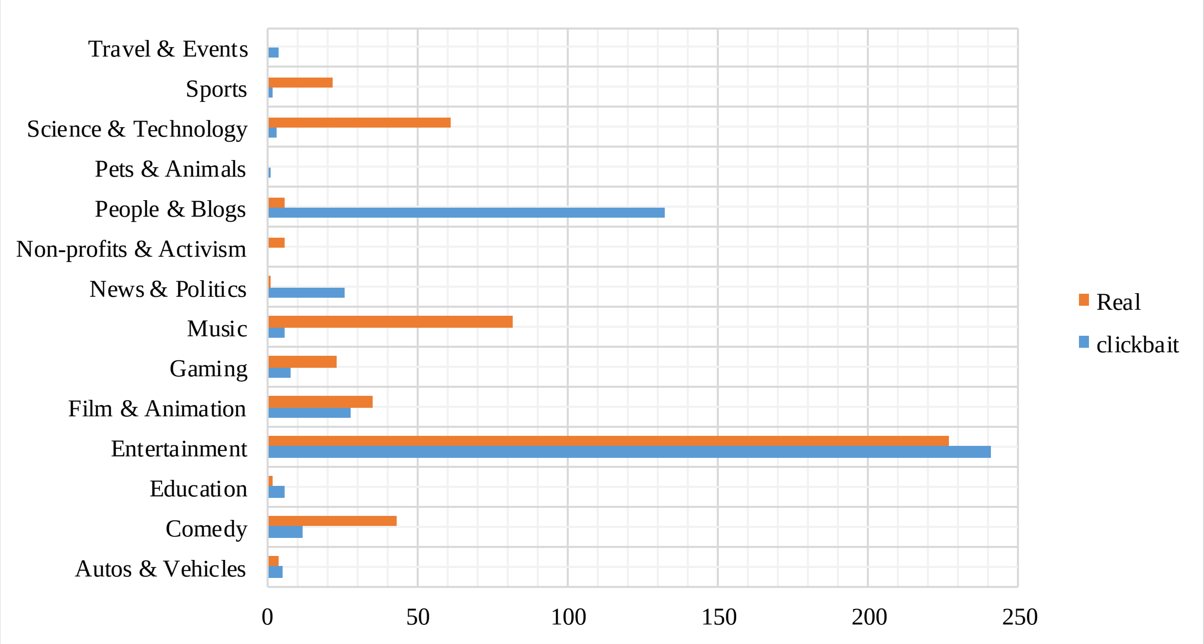


Fig.4.2.4. Number of Videos by Category and Class

[33] https://www.youtube.com/channel/UC_UdS7tWCwgBDoaM-Hmkzxg/videos

Table 4.2.1. Detailed Description of the Self-Generated Dataset (MVD)

S.NO	Category	Number of Videos	Class
1	Autos & Vehicles	5	clickbait
2	Autos & Vehicles	4	Non-Clickbait
3	Comedy	12	clickbait
4	Comedy	43	Non-Clickbait
5	Education	6	clickbait
6	Education	2	Non-Clickbait
7	Entertainment	241	clickbait
8	Entertainment	227	Non-Clickbait
9	Film & Animation	28	clickbait
10	Film & Animation	35	Non-Clickbait
11	Gaming	8	clickbait
12	Gaming	23	Non-Clickbait
14	Music	6	clickbait
15	Music	82	Non-Clickbait
16	News & Politics	26	clickbait
17	News & Politics	1	Non-Clickbait
18	Non-profits & Activism	6	Non-Clickbait
19	People & Blogs	132	clickbait
20	People & Blogs	6	Non-Clickbait
21	Pets & Animals	1	clickbait
22	Science & Technology	3	clickbait
23	Science & Technology	61	Non-Clickbait
24	Sports	2	clickbait
25	Sports	22	Non-Clickbait
26	Travel & Events	4	clickbait
Total		**474 clickbait** **513 Non-Clickbait**	

4.2.2 Problem Definition

We define the clickbait detection task as for a given set of videos, the system has to determine which of the videos are reporting clickbait's and that does not faithfully represent the event it refers to. The identification of clickbait videos is ultimately meant to warn users that the given video content does not faithful about the claim it represents, and helps in countering the spreading of false content. In this, we have considered the following three detection cases shown below in Table 4.2.2.

Table 4.2.2. Possible cases for the clickbait's detection

S.NO	Detection Cases
1.	**The title** faithfully represents the **video speech content** and **comments** are not in contradiction
2.	**The title** does not faithfully represent the **video speech content** and both are in contradiction.
3.	**The title** faithfully represents the **video speech content** and **comments** are in contradiction.

A set of evidence is required to justify and verify the above cases and warn the user to think twice while believing and spreading false information. If these three cases are identified, then there is a possibility that the video is clickbait and does not faithfully represent the event that it refers to. Formally, the task takes a set of video ids $V_{ID} = V_{ID1}, V_{ID2}, V_{ID3} \dots . V_{IDN}$ as an input, and the classifier has to determine whether each of these videos V_{IDi} is a clickbait or Non-Clickbait by assigning a label from $Y = \{C, R\}$. Hence, we formulate the task as a binary classification problem, whose performance is analyzed and evaluated by computing the various performance measures like precision, recall, and F1 score for the target class, i.e., Clickbait. There are three types of cognitive evidence that have been considered for the detection of clickbait. Each of these evidence gives a significant contribution in finding the clues in predicting a video as clickbait or not. The three sets of cognitive evidence for a video are as follows.

4.2.3 Detecting Clickbait's Videos

In this section, we describe the proposed method, the CVD (Clickbait Video Detector) to address the problem formulated previously. The technique consists of three major pieces of evidence, retrieved using three feature components based on Video, Human-consensus, and User-Profile. The first feature component is used to extract video-related features (e.g. speech-title similarity, number of likes, number of dislikes, dislike-like ratio). The second feature component is based on human-consensus. This module learns from individual human cognition and combined it with the consensus response. The output has been retrieved, which gives the agreement of individuals towards the posted content. The third component is the user-profile feature extraction (e.g. video-age-ratio, channel views, registration age). This is directly related to the reputation of the video uploader. Lastly, we finish with the classification model, which finally does the binary classification (clickbait and Non-Clickbait) using features extracted from the first three components. The overview of CVD is shown in Fig.4.2.5, where the three sets of features are extracted from the video, comments, and channel information.

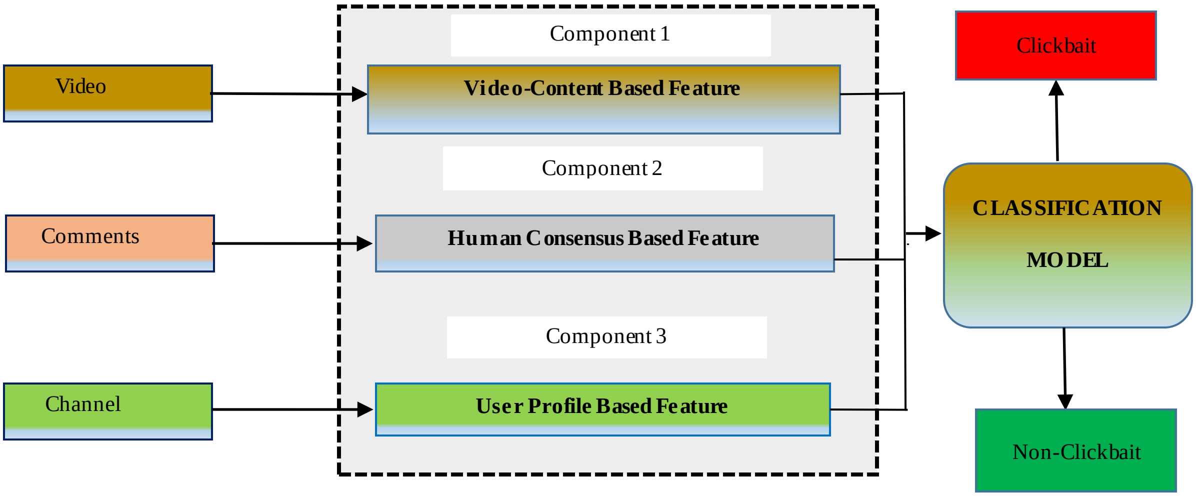


Fig.4.2.5. An Overview of CVD

- **Video-based features**

Video-Content based feature is the first component of our proposed method. This component is responsible for the extraction of video-content-based features (e.g. speech-title similarity, number of likes, number of dislikes, dislike-like ratio). The speech-title similarity is one of the crucial features and plays a major role in retrieving Evidence 1. The speech-title similarity is the similarity of speech text with respect to the title of the video, which identifies whether the given claim attached to the video, faithfully represents the event that it refers to. To identify how faithfully the video is representing a claim, speech has been extracted from each video. The Google speech to text API has been used to the speech part, that later be converted into text. The cosine similarity has been applied in between the text extracted from the speech part of the video and the title, to measure the similarity among them. Google's speech to text API has been used for speech recognition. The speech to text has three main methods to perform speech recognition (Synchronous, Asynchronous, and Streaming Recognition)[34]. Here we have used the synchronous recognition method, as in our case we need to process the data of less than one 1 minute and synchronous recognition requests are limited to audio data of 1 minute or less in duration. The detailed description of how the complete process is followed is shown

[34] https://cloud.google.com/speech-to-text/docs/basics

in Fig.4.2.6. It shows the retrieving process of the speech-title similarity for Evidence 1. In the first step, the video is given as an input, which is translated into its audio format. The audio is converted into text format using Google Speech API. So, to get better similarity results, the audio is processed in parts. For each video, the 1 min segments have been analyzed, as we are getting enough information within this duration to predict a video is bogus or credible. The audio transcripts of 1 min are subdivided into 4 parts of each 15 secs. The four text parts that are incorporated in Fig.4.2.6 is to split the 1-minute audio transcripts into smaller subparts to identify how similar the title/claim with respect to the content that is presented in the video. It has been noticed that in YouTube videos the given titles are too short in length, containing very few words and a 1-minute audio transcript considered for analysis is quite lengthy and is not so effective to apply cosine similarity between them for reliable prediction. To address this problem, the audio transcript is split into small subpart (4 text parts of 15 secs each) (Text part 1, Text part 2, Text part 3, and Text part 4), each text subpart contains 15-sec audio transcripts in sequence after which the individual text similarity with respect to the title has been calculated, later the average value has been considered as the final similarity value. The video-content-based features are shown in Table 4.2.3. These features are evaluated to represent the statistical characteristics of the video content.

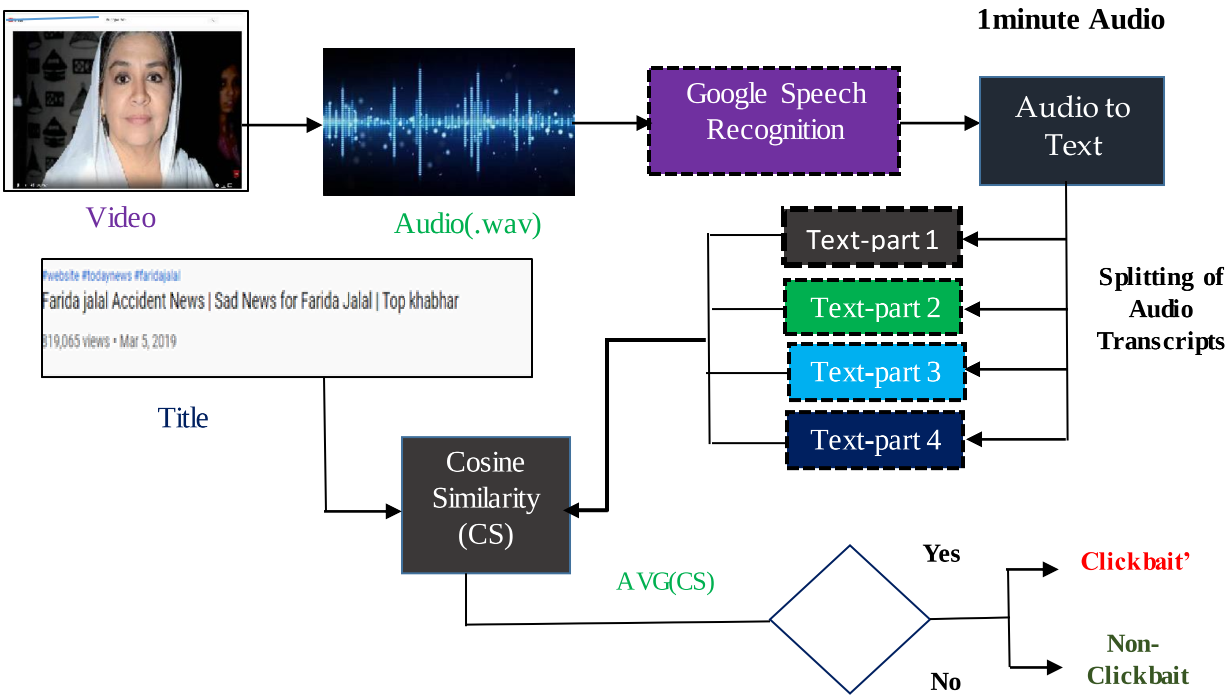

1minute Audio
Video
Audio(.wav)
Google Speech Recognition
Audio to Text
Text-part 1
Text-part 2
Text-part 3
Text-part 4
Splitting of Audio Transcripts
#website #todaynews #faridajalal
Farida jalal Accident News | Sad News for Farida Jalal | Top khabhar
819,065 views • Mar 5, 2019
Title
Cosine Similarity (CS)
AVG(CS)
Yes
No
Clickbait'
Non-Clickbait

Fig.4.2.6. The Figure represents the process of retrieving Evidence 1.

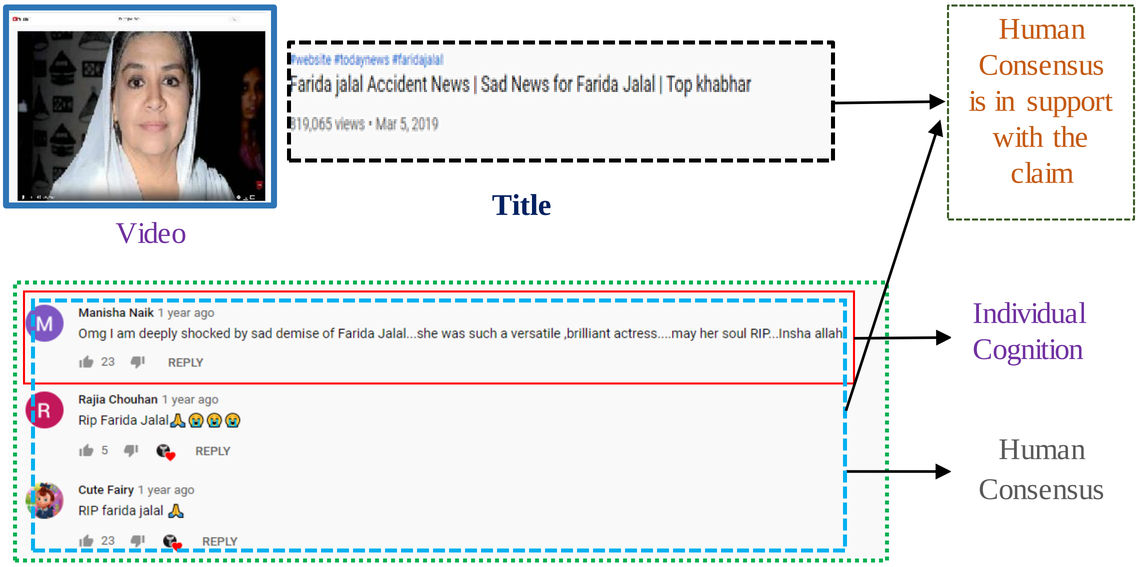
Table 4.2.3. Video-Content based feature

Feature	Description
Audio Transcript based Features (Avg_cs)	The average cosine similarity measure between audio transcripts and the title of the video. This is one of the novel features and very few studies incorporate it.
Number of Likes $L(x)$:	This feature represents the number of likes on a video.
Number of Dislikes $D(x)$	This feature represents the number of dislikes on a video.
Dislike to Like Ratio $DL(x)$	The ratio of the number of dislikes to like count on a video. It has been observed that clickbait videos received more dislikes compared to likes. $$\boldsymbol{DL(x)} = \frac{D(x)}{L(x)}$$
Number of Views	The number of views Received by a video.

- **Human Consensus-Based Feature**

Individual human cognition can play an important role and gives a significant contribution in forming evidence for the detection of Clickbait's and the second component of our proposed approach. The individual viewer has their own cognition that has been come out in the form of expression/ emotions given as a video response. Many of the malicious users have not allowed the comments on their video because the human consensus gives an initial clue about a video, and if any new user visits the page, they can make an initial thought about video credibility via reading how individual responding to a video. That's why most of the time, it can be seen that commenting is not enabled on videos created with some malicious intent. Fig.4.2.7 shows the process of retrieving Evidence 2. Here multilingual content has also been addressed. If any of the content is found to be in a different language, it is translated into English text using google translator, then further be used for analysis. So, here we have also addressed the multilingual content.

A total of 6 Human consensus-based features are extracted. The Human-consensus based features are shown in Table 4.2.4. These features are evaluated to represent the statistical characteristics of the responses of the viewers.



#website #todaynews #faridajalal

Farida jalal Accident News | Sad News for Farida Jalal | Top khabhar

319,065 views • Mar 5, 2019

Manisha Naik 1 year ago

Omg I am deeply shocked by sad demise of Farida Jalal...she was such a versatile ,brilliant actress....may her soul RIP...Insha allah

23 REPLY

Rajia Chouhan 1 year ago

Rip Farida Jalal🙏😭😭😭

5 REPLY

Cute Fairy 1 year ago

RIP farida jalal 🙏

23 REPLY

Fig.4.2.7. The Figure represents the process of retrieving Evidence 2.

Table 4.2.4. Human-Consensus based feature

Feature	Description
Number of Comment c(x)	This feature represents the number of comments received on a video. To restrict our search analysis, in total maximum of 200 comments have been considered. The below equation represent the comment count. $\boldsymbol{c}(\boldsymbol{x}) = \sum_{i=1}^{200}(\text{c i })$
Positive Polarity p(x)	This is the feature that indicates, how many comments showing a positive opinion towards a video. $\boldsymbol{p}(\boldsymbol{x}) = positive\ polarity\ count\ /c(x)$
Negative Polarity $\boldsymbol{n}(\boldsymbol{x})$	This is the feature that indicates, how many comments showing a negative opinion towards a video. $\boldsymbol{n}(\boldsymbol{x}) = negative\ polarity\ count/ c(x)$
Positive-Negative Polarity Ratio $\boldsymbol{pn}(\boldsymbol{x})$	This is the ratio of positive to negative comment polarity count. $\boldsymbol{pn}(\boldsymbol{x}) = \frac{p(x)}{n(x)}$
Fake_Comment_Count $\boldsymbol{FCC}(\boldsymbol{x})$	The fake comment count is the number of comments having clickbait phrases. **Clickbait's Phrases(CP)=** {fake, bullshit, hoax, wrong… etc.} $\boldsymbol{FCC}(\boldsymbol{x}) = \sum_{i=1}^{200}(\text{CP })$
Fake_Comment_Count Ratio $\boldsymbol{FCCR}(\boldsymbol{x})$	It is the ratio of the number of fake_comment_count to the total number of comments encountered. $\boldsymbol{FCCR}(\boldsymbol{x}) = FCC/c(x)$

- **User-Profile based Features**

The reputation of the individual channel also plays an important role in identifying the credibility of the uploaded video. (e.g. video-age-ratio, channel views, registration age). A total of 7 User-Profile-based features are extracted. The User-Profile based features are shown in Table 4.2.5. These features are evaluated to represent the statistical characteristics of the responses of the viewers.

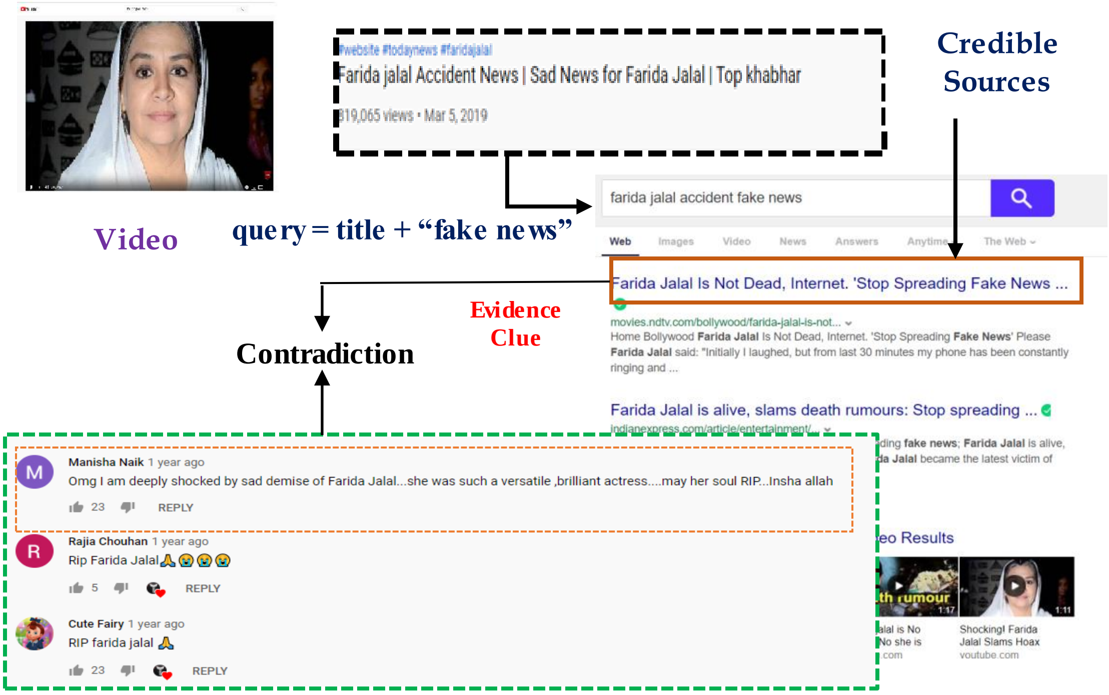
Table 4.2.5. User Profile-based Features

Feature	Description
Registration Age $\boldsymbol{r(x)}$	The age of the user is an indicative measure of the rounded number of days that the user has spent on YouTube, i.e. from the day account was created up to the day of the current post.
Channel Views $\boldsymbol{CV(x)}$	The total number of views received by the channel.
Total_no_of_Videos $\boldsymbol{V(x)}$	The total number of videos has been posted by the channel till date.
Subscriber Count $\boldsymbol{SC(x)}$	The total number of subscribers count on the channel.
Video_to_Age_ratio $\boldsymbol{VA(x)}$**:**	This is the ratio of the total number of videos uploaded by the channel to its registration age. $\boldsymbol{VA(x)} = \frac{V(x)}{r(x)}$
Subscribers_to_Age_ratio $\boldsymbol{SA(x)}$	This feature represents the ratio of the number of subscribers on the channel to its registration age. $\boldsymbol{SA(x)} = \frac{SC(x)}{r(x)}$
Channel_Views_to_Subscribers $\boldsymbol{CS(x)}$	It is a ratio of the number of views received by the channel to its subscriber count. $\boldsymbol{CS(x)} = \frac{CV(x)}{r(x)}$

From the all given set of features there are some important findings, it has been observed that the Non-Clickbait video channel has more subscribers with respect to registration age as compared to clickbait's video channel. The findings reveal that the average cosine similarity for Non-Clickbait videos is too less than that of the clickbait's, one of the reasons may be because the clickbait videos probably repeat the same sentence as mentioned in the title, of many times to extend the length of the video with redundant and bogus content. It has also been found that the number of dislikes is more than the number of likes for many of the videos belongs to clickbait's as compared to non-clickbait's.

- **Credible Sources**

Credible sources are the third sort of evidence, which plays an important role in news verification. The two sets of evidence that we have discussed earlier give significant information, however, they fail in certain situations. The first situation is when the speech-title similarity perfectly matches as well as comments are also in support. So, are these measures sufficient enough? It may happen that, these two pieces of evidence are in support, even though the information is false as shown in Fig.4.2.8 , it clearly shows that evidence 1 and evidence 2 are in support, but the information is false. So what is the breaking point here, through which we can reliably conclude about the credibility of news?. None other but the credible news sources, we need to scrap the news headlines related to the specified claim by searching it on google, and searching for clickbait phrases in the headlines like {*'misleading', 'misinformation', 'not known', 'no proof', 'no known', 'no scientific evidence', 'no evidence', 'not verified',' hoax', 'clickbait', 'not proven', 'denied', 'deny', 'unverified', 'false', 'fake', 'fake news', 'falsely', 'myth', 'ridiculous', 'rumour',' not dead',' death rumours'*} for justifying the claim. There can be another case of it where it can be applicable, when the uploader is not allowed any comments and make the commenting off, where it is quite needful to retrieve Evidence 3. The query is build using the video title concatenated with the fake news keywords *query= "title+ fake news"*, which goes as a search query to google. The top 15 URLs concerning the specific claim are scrapped and analyzed. These 15 web titles are considered as a replica of video comments, and the same measures are identified here, as evaluated over video comments (Human-consensus based features) to get the informative clues when comments are not available.



#website #todaynews #faridajalal
Farida jalal Accident News | Sad News for Farida Jalal | Top khabhar
319,065 views • Mar 5, 2019

farida jalal accident fake news

Farida Jalal Is Not Dead, Internet. 'Stop Spreading Fake News ...
movies.ndtv.com/bollywood/farida-jalal-is-not...
Home Bollywood **Farida Jalal** Is Not Dead, Internet. 'Stop Spreading **Fake News**' Please **Farida Jalal** said: "Initially I laughed, but from last 30 minutes my phone has been constantly ringing and ...

Farida Jalal is alive, slams death rumours: Stop spreading ...
indianexpress.com/article/entertainment/...
ding **fake news**; **Farida Jalal** is alive,
da Jalal became the latest victim of

eo Results
th rumour 1:17
alal is No
No she is
.com
1:11
Shocking! Farida Jalal Slams Hoax
youtube.com

Manisha Naik 1 year ago
Omg I am deeply shocked by sad demise of Farida Jalal...she was such a versatile ,brilliant actress....may her soul RIP...Insha allah
23 REPLY

Rajia Chouhan 1 year ago
Rip Farida Jalal🙏😭😭😭
5 REPLY

Cute Fairy 1 year ago
RIP farida jalal 🙏
23 REPLY

Fig.4.2.8 The Figure represents the process of retrieving Evidence 3.

Algorithm 1 gives the detailed procedure of the retrieval process of all three pieces of evidence, and from all sets of features, some of the features that play a crucial role in retrieving the evidence are identified. The given algorithm takes a set of videos as an input and returns the status that whether a video is clickbait or not as an output, the given algorithm will further be used to create a Non-Clickbait-time prediction of clickbait. The crucial video-based features encountered are Average cosine similarity (Avg_cs) and Dislike-like ratio (DL) for retrieving Evidence 1, and from the analysis of all sets of sample the threshold values have been identified, if the Avg_cs >0.10 OR DL>=0.40 then the video is likely to be clickbait concerning the video content and make the value of Evidence 1 to be true or 1, otherwise 0. On the other hand, FCCR (Fake Comment Ratio) plays an important role in retrieving Evidence 2. After getting the value of Evidence 1 and Evidence 2, three different defined cases have been observed concerning these values and the output status has been predicted as Clickbait or Non-Clickbait.

Table 4.2.6. The table shows the cases and the needful evidence required for the detection.

S.NO	Detection Cases	Essential Measure	Desirable Measure
1.	**Title** faithfully represents the **video speech content** and **comments** are not in contradiction	Evidence 1 Evidence 2 Evidence 3	Evidence 1 Evidence 2 Evidence 3
2.	**Title** does not faithfully represent the **video speech content** and both are in contradiction.	Evidence 1	Evidence 1 Evidence 2
3.	**Title** faithfully represents the **video speech content** and **comments** are in contradiction.	Evidence 1 Evidence 2	Evidence 1 Evidence 2

Algorithm 1.(Clickbait Video Detection)

Input(Video_id) and Output(Status)

```
def. func1(Video_id)
    Evidence1= Video_based_feature ();
    Evidence2= Human_consensus_based_feature ();
    If (Evidence1==0) AND (Evidence2==0):
        Scrapped_urls= Google_search(query)
          Evidence 3= Processing(Scrapped_urls)
          IF (Evidence 3==1):
              Status= Print("Clickbait")
          Else:
              Status= Print("Non-Clickbait")
     Elif (Evidence1==1 and Evidence2==1):
         Status = Print("Clickbait")
def. Video_based_feature ()
   Evidence1 = 0
   Title= Extract_Title ();
   Audio_transcript= Extract_Transcript ();
   number_of_dislike= Count(Dislikes);
   number_of_like= Count(like);
  Averge_cosine_similarity(Avg_cs) = Cosine_similarity (Title, Audio_transcript);
  Dislike_like_ratio(DL)= (number_of_dislike)/ (number_of_like)
     If (Avg_cs>0.10 OR DL>=0.40):
         Evidence1= 1
         return (Evidence 1)
    Else:
         Evidence1=0
         return (Evidence 1)
def. Human_consensus_based_feature ()
    Evidence2= 0
    $Number\ of\ comment\ c(x) = \sum_{i=1}^{200}(\mathbf{c\ i}\ )$
    Fake_Comment_Count_Ratio(FCCR)= Fake_Comment_Count /c(x);
    If(FCCR>=0.015)
      Evidence2= 1
      return(Evidence2)
   Else:
       Evidence2= 0
       return(Evidence2)
```

4.2.4 Experimental Analysis and Results

In this section, we evaluate the performance of the CVD scheme in comparison with the state-of-the-art method. The result shows that the CVD technique significantly outperforms the baseline method with respect to different performance measures.

Table 4.2.7 and Table 4.2.8 briefly describe the results for different classifiers using all sets of features on the Self-Generated Dataset (MVD) by employing two validation strategies, Cross-validation, and Percentage Split, respectively. From the rigorous analysis of all the proposed features, it has been observed that Human Consensus and User Profile features give a significant contribution and play a major role in predicting the video as clickbait or non-clickbait. Table 4.2.7 shows the performance analysis of the model by employing a cross-validation technique concerning different measures, TP (True Positive), False Positive(FP), PRE(Precision), REC(Recall), FM(F-Measure), and ACC(Accuracy). The analysis has been presented by considering all sets of features, video-based features, human consensus-based features, and user-profile features. The performance of the classifiers (Random-forest, Naïve-Bayes, Logistic, SVM, SGD, k-nearest, and J48 using all sets of features and for each independent set of features suggests that J48 remarkably outperforms over the rest of the classifiers with the highest accuracy of 98.28 on both cross-validation and percentage-split mechanism. This can be seen clearly when we look at precision, where J48 performs substantially better than the rest. However, the k-nearest neighbor classifier performs better, only when considering user-profile features. It has been observed that SVM is not performing well when considering only video-based features, while significant improvement in the accuracy, when considering all sets of features. Whereas, logistic regression performs worst in comparison to all other classifiers in each scenario. On the other hand, using the percentage split technique, the random-forest, and J48 both performing the same on all sets of features in terms of true positive, precision, recall, f-measure, and accuracy.

Table 4.2.7: Performance of the various classifier by employing Cross-Validation.

All Set of Features								
Classifiers	Fold1	Fold2	TP	FP	PRE	REC	FM	ACC
Random Forest	10	-	0.974	0.025	0.975	0.974	0.974	97.37
Naïve Bayes	10	-	0.964	0.034	0.965	0.964	0.964	96.37
Logistic	10	-	0.909	0.085	0.922	0.909	0.909	90.92
SVM	10	-	0.955	0.042	0.959	0.955	0.955	95.46
SGD	10	-	0.955	0.043	0.957	0.955	0.955	95.46
k-nearest	10	-	0.964	0.037	0.964	0.964	0.964	96.37

J48	10	-	**0.983**	0.017	**0.983**	0.983	0.983	**98.28**
Random Forest	-	20	0.974	0.025	0.974	0.974	0.974	97.37
Naïve Bayes	-	20	0.965	0.033	0.966	0.965	0.965	96.47
Logistic	-	20	0.888	0.105	0.905	0.888	0.887	88.81
SVM	-	20	0.955	0.042	0.959	0.955	0.955	95.46
SGD	-	20	0.957	0.041	0.959	0.957	0.957	95.66
k-nearest	-	20	0.965	0.035	0.965	0.965	0.965	96.47
J48	-	20	**0.983**	0.017	**0.983**	0.983	0.983	**98.28**
Video-based features								
Random Forest	10	-	0.875	0.120	0.885	0.875	0.875	87.5
Naïve Bayes	10	-	0.772	0.243	0.830	0.772	0.760	77.21
Logistic	10	-	0.755	0.255	0.773	0.755	0.749	75.50
SVM	10	-	0.542	0.490	0.757	0.542	0.406	54.23
SGD	10	-	0.746	0.264	0.763	0.746	0.740	74.59
k-nearest	10	-	0.876	0.125	0.876	0.876	0.876	87.60
J48	10	-	**0.898**	0.102	**0.898**	0.898	0.898	**89.81**
Human Consensus-based Feature								
Random Forest	10	-	0.884	0.124	0.904	0.884	0.882	88.40
Naïve Bayes	10	-	0.856	0.154	0.883	0.856	0.852	85.58
Logistic	10	-	0.795	0.218	0.845	0.795	0.786	79.53
SGD	10	-	0.823	0.190	0.868	0.823	0.816	82.25
k-nearest	10	-	0.891	0.111	0.893	0.891	0.891	89.11
J48	10	-	**0.902**	0.104	**0.913**	0.902	0.901	**90.22**
User Profile-based features								
Random Forest	10	-	0.955	0.042	0.959	0.955	0.955	95.46
Naïve Bayes	10	-	0.949	0.049	0.952	0.949	0.949	94.85
Logistic	10	-	0.800	0.187	0.857	0.800	0.794	80.04
SVM	10	-	0.955	0.042	0.959	0.955	0.955	95.46
SGD	10	-	0.949	0.048	0.953	0.949	0.949	94.85
k-nearest	10	-	**0.973**	0.027	**0.973**	0.973	0.973	**97.27**
J48	10	-	0.971	0.030	0.971	0.971	0.971	97.07

Fig.4.2.9. shows the comparative analysis of various classifiers (Random-Forest, Naïve-Bayes, Logistic, SVM, SGD, k-nearest, and J48) on a different set of features by applying 10-fold cross-validation. The comparison results in terms of accuracy measure, clearly show that the model outperforms when employing all three features combinedly, instead of applying individual features. However, it can also be observed that User-profile features individually perform better in comparison to Human-consensus and Video-content based features. Whereas Video-content based features do not perform well individually. The experimental results reveal that the user-profile-based features significantly improve the overall performance of the proposed model compared to other feature sets. One of the main reasons identified from the observation is that the reputation of an individual channel/account/user profile plays an

important role in identifying the credibility of the uploaded video and gives an efficient clue for the verification of misleading content[47][14]. Previous studies also reveal that user profile-based features are efficient in detecting false content.

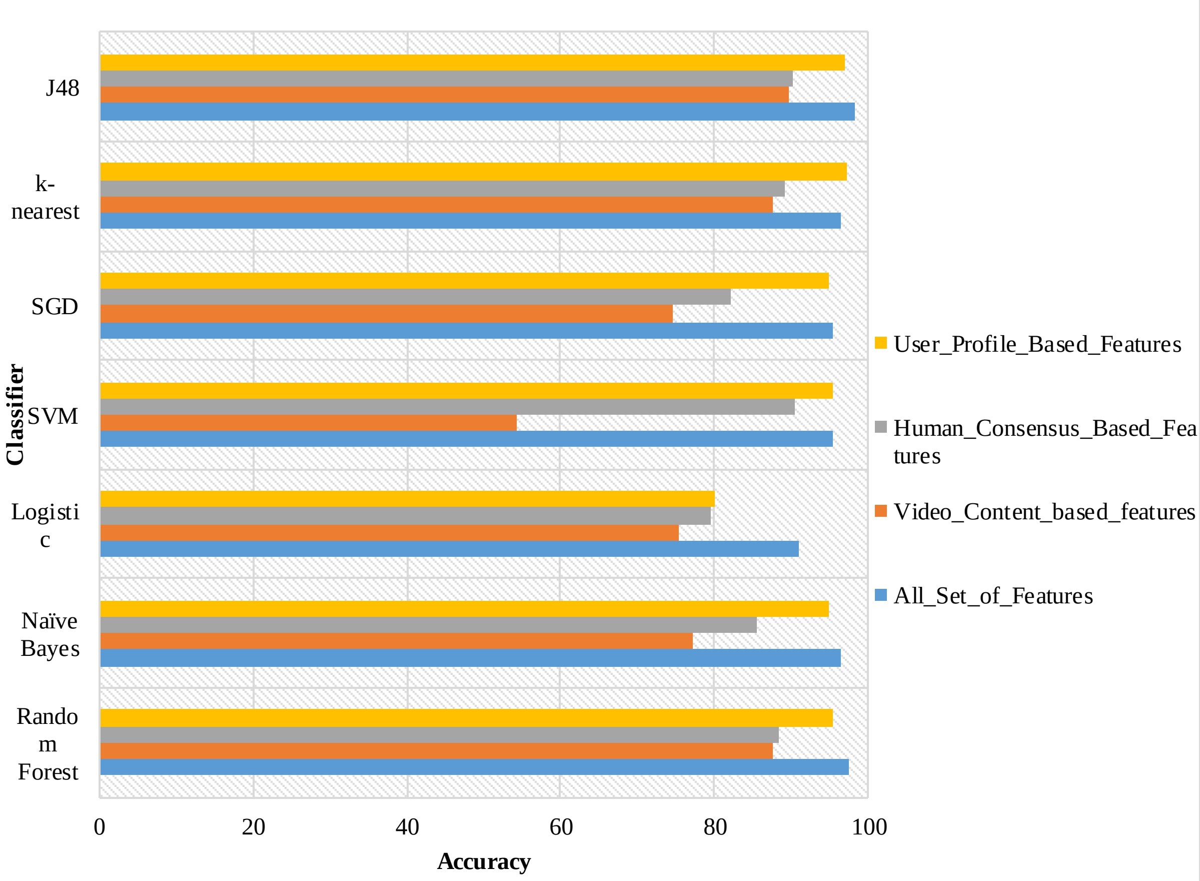


Fig.4.2.9 Comparative Analysis of various classifiers on different set of features.

The authors of [47][14][101][102][103][104], reports that the user profile/ account-based features plays a significant role in detecting false information. Whereas, the other two features are not performing well compared to these features due to some constraints like the video-content based features are relying on the similarity of audio transcripts and the title of the video but there are some cases where the audio is too noisy due to which the clear transcripts can't be retrieved for matching or in a case when the video doesn't have any speech content present. These cases may be liable to degrade the performance of the model. Whereas, human-consensus-based features are also performing well after user profile features. However, in some cases when sufficient clickbait phrases are not matched/ identified from the user responses or credible link sources, in that case, this feature may lack in performance.

Table 4.2.8: Performance of the various classifier by employing Percentage Split.

Classifiers	Split1	Split2	TP	FP	PRE	REC	FM	ACC
Random Forest	70:30	-	0.977	0.024	0.977	0.977	0.977	**97.65**
Naïve Bayes	70:30	-	0.963	0.037	0.964	0.963	0.963	96.30
Logistic	70:30	-	0.903	0.098	0.907	0.903	0.902	90.26
SVM	70:30	-	0.899	0.102	0.916	0.899	0.898	89.93
SGD	70:30	-	0.956	0.044	0.957	0.956	0.956	95.63
K-nearest	70:30	-	0.970	0.030	0.970	0.970	0.970	96.97
J48	70:30	-	**0.977**	0.023	0.977	0.977	0.977	**97.65**
Random Forest	-	80:20	0.975	0.026	0.975	0.975	0.975	**97.47**
Naïve Bayes	-	80:20	0.929	0.067	0.936	0.929	0.929	92.92
Logistic	-	80:20	0.899	0.105	0.907	0.899	0.898	89.89
SVM	-	80:20	0.949	0.054	0.954	0.949	0.949	94.94
SGD	-	80:20	0.960	0.042	0.961	0.960	0.960	95.95
k-nearest	-	80:20	0.970	0.030	0.970	0.970	0.970	96.96
J48	-	80:20	**0.975**	0.025	0.975	0.975	0.975	**97.47**

The AUC-ROC curve using the Random Forest, Naïve-Bayes, and K-nearest neighbour classifier model is shown in Fig.38. To get a good understanding of the performance of the model, we can also look at their receiver operating characteristics, ROC curves. Fig.38. represents the three ROC curves for the random forest, Naïve Bayes, and K-Nearest Neighbour classifier, trained on all sets of features (video-based, human-consensus, and user-profile based). Here we can see that the peak value for the random forest is achieved at the point (x=0.01129, y=1) with having a minimum false positive rate of 0.01. At this point, the model would correctly identify 100% of the true rumors, with only getting around 1% of the false rumors mistakenly classified as true. Whereas, for Naïve Bayes the peak value is achieved at the point (x=0.401, y=1), having a minimum false positive rate of 0.406. At this point, the model would correctly identify 100% of the true rumors, with only getting around 40% of the false rumors mistakenly classified as true. Depending upon the application purpose, the user can choose or pick up different points on the ROC curve, like with respect to the normal user, who want to find the truthfulness of the specific news, the user could perhaps choose the point on the ROC curve with having minimum false positive rate, or the point where FPR is closer to zero, for getting reliable information.

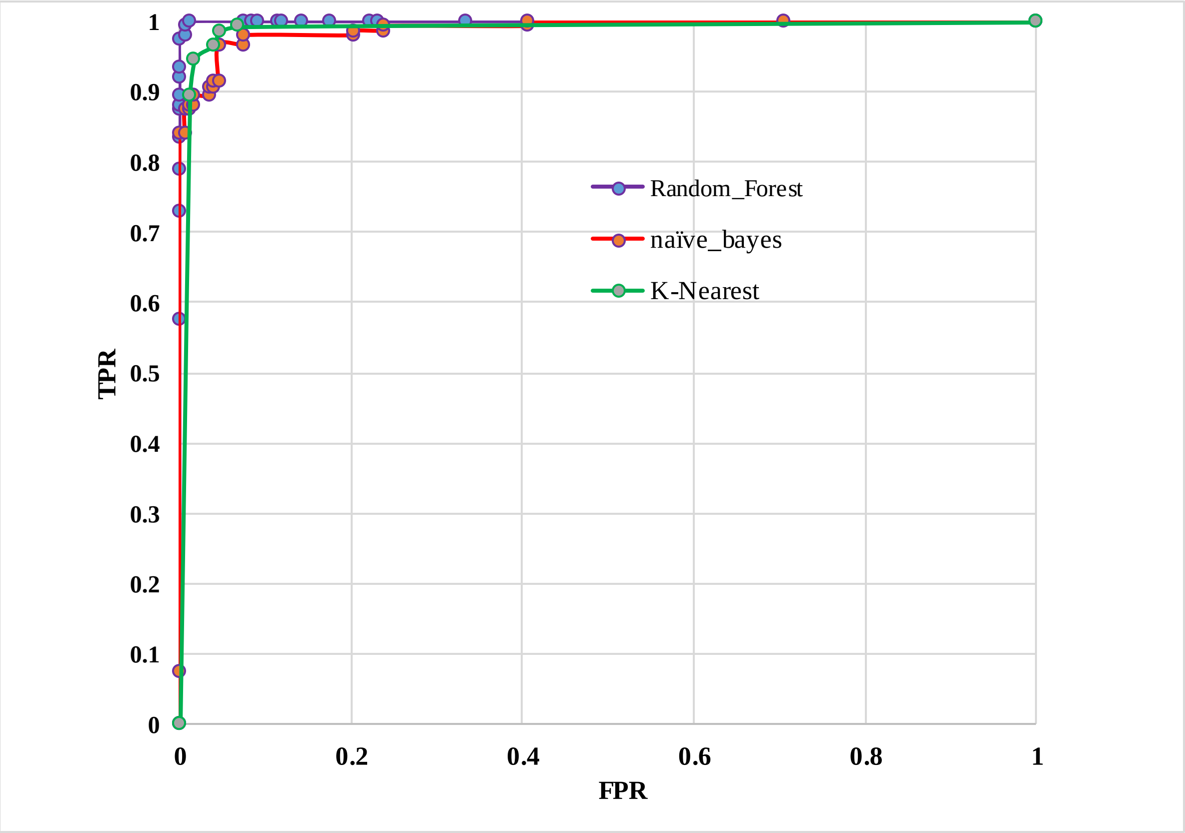


Fig.4.2.10 The AUC-ROC Curve

Table 4.2.9: Description of features used for the plot of Scatter matrix representation as shown in Fig.4.2.11.

Feature No	Feature	Feature No	Feature
X1	DL(x)	Y1	CV(x)
X2	CS(x)	Y2	FCCR(x)
X3	Avg_cs	Y3	VA(x)
X4	Video_id	Y4	SA(x)
X5	PN(x)	Y5	PN(x)
X6	SA(x)	Y6	Video_id
X7	VA(x)	Y7	Avg_cs
X8	FCCR(x)	Y8	CS(x)
X9	CV(x)	Y9	DL(x)

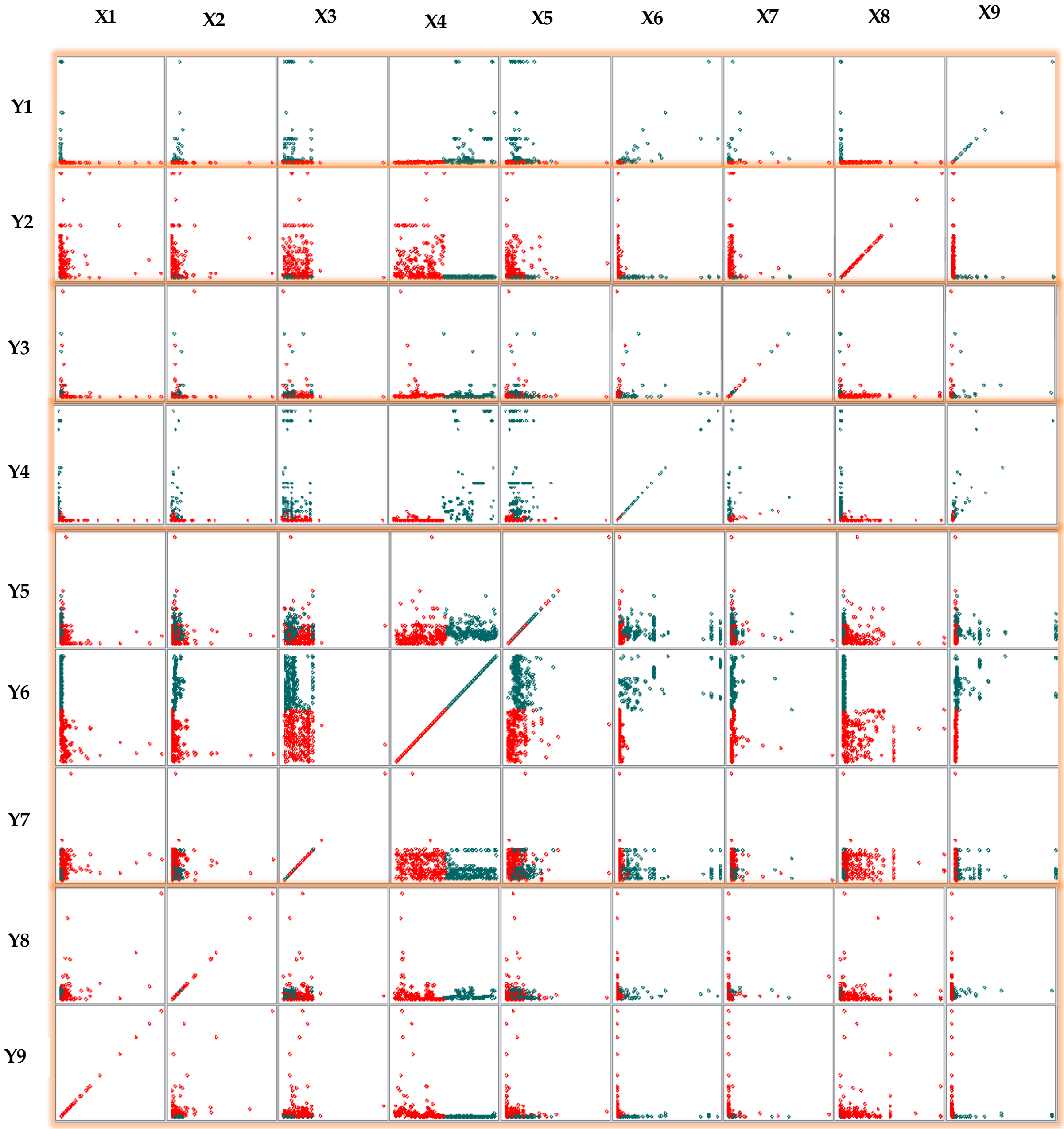


Fig.4.2.11 Plot Matrix representation of proposed features against all other features for Clickbait's (Red) and Non-Clickbait's (Green).

Along with this, the Scatter plot matrix representation of the features against other features for clickbait and Non-Clickbait data sample is also represented to give a visual qualitative understanding of the correlation. We have created the scatter plot of one feature against another as shown in Fig.4.2.11. Table 4.2.9 represents the features on the X and Y-axis of the plot. The plot visually represent the relationship between each of the feature on the set $X = X_1, X_2 \ldots. X_9$ on X-axis to the set $Y = Y_1, Y_2 \ldots. Y_9$ on Y-axis. The representation is useful, as it is showing the pattern in the relationship between attributes to visually explore the relationship between several numeric values. The dots in the scatter plot are colored by their class value (Clickbait and Non-Clickbait)[35]. Like, it can be seen that the scatter plot matrix of subscriber_age_ratio feature on the X-axis against all other features shows an approximately clear separation between two classes (Clickbait's and Non-Clickbait's) and shows how the points are correlated with respect to different classes.

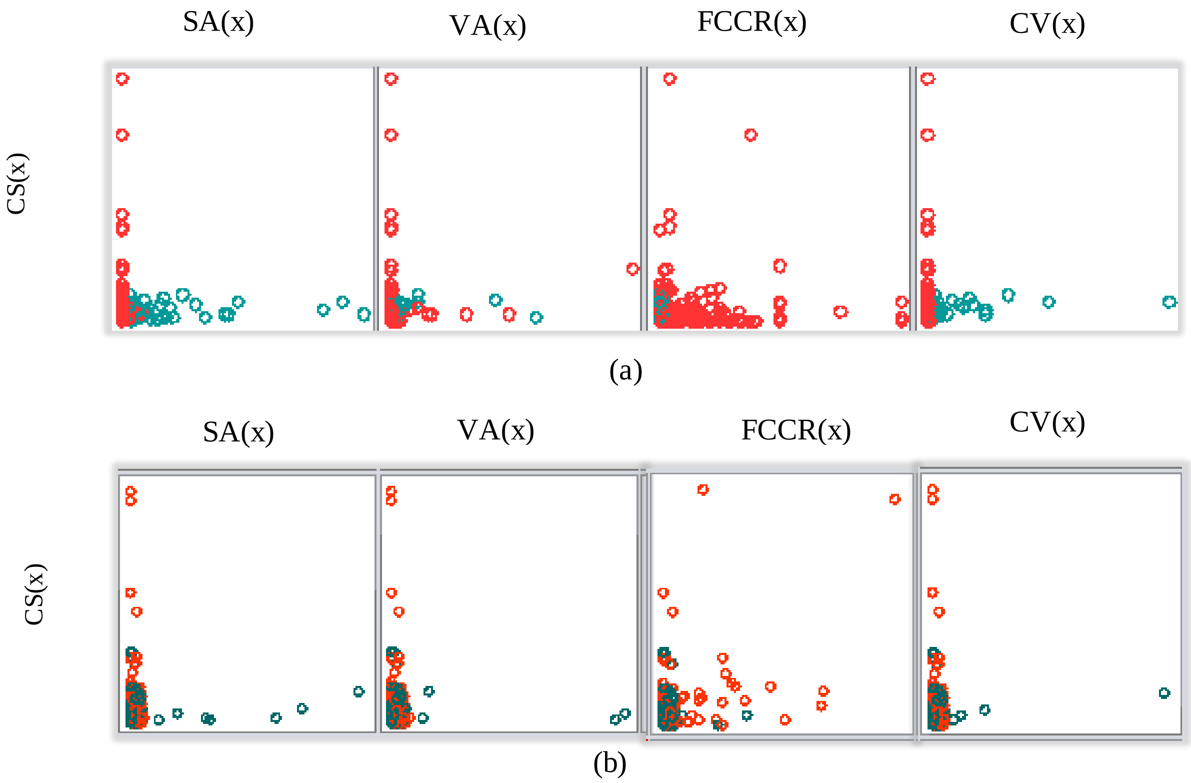
Along with the self-generated dataset, as we discussed previously, the analysis has also been applied over the other state-of-the-art methods. It has been observed that some of the work contributed by creating a public dataset of fake/misleading videos [98],[97], [105]. However, still very few datasets are available for comparative analysis, there is a small, but some of the datasets of fake videos on YouTube are publicly available called FVC (Fake Video Corpus)[95] and FVC- 2018[105]. The dataset[36] is the collection of 381 videos in which 201 are fake and 180 are Non-Clickbait. After analysis, it has been found that most of the videos are removed from YouTube. Due to this, we are only able to crawl 84 fake and 90 Non-Clickbait videos, we divide these videos into two disjoint sets, FVC (70:30), with having 70 videos for training and 30 videos for testing. The comparative analysis with the state-of-the-art on the same dataset is shown in Table 4.2.10. Whereas, the other dataset that is considered for analysis is FVC 2018. The FVC-2018 is the extended version of the FVC dataset i.e, the samples in the FVC-2018 is an order of magnitude larger than that of FVC, and much more varied. The dataset was extended with 3,729 additional fake videos and 2,283 real videos, published on YouTube, Facebook or twitter, and considering the time period of April 2006 and June 2018. However for analysis purpose we have considered only YouTube(YT) Videos i.e 1,675(Fake) and 993(Real), comparison analysis has been performed on the same[37]. From the previous studies, it has been found that a very limited number of baselines are available for clickbait's video detection[100]. The online clickbait video detection problem is an emerging field and is a

[35] https://machinelearningmastery.com/better-understand-machine-learning-data-weka/
[36] https://github.com/MKLab-ITI/fake-video-corpus/blob/master/FVC.csv
[37] https://github.com/MKLab-ITI/fake-video-corpus/blob/master/FVC_dup.csv

largely unsolved research problem. Due to which very few works have been reported yet. Along with this, very limited datasets are publicly available for the evaluation of the proposed algorithm. Many of the works have not released the source code as well. Due to all these research constraints, limited baselines are available for comparative analysis. Some of the prominent methods that are closely related to our work are described in the following paragraph.

The authors of [105], proposed a verification algorithm to detect fake videos. They have also created the FVC-2018 dataset to train and evaluate the proposed method. In the verification algorithm, the author applied the same process that was used in [98], along with it two model variants: a concatenation of the two feature sets(videos metadata and comments feature) and the agreement-based approach given in [79] was used. Their proposed algorithm has been evaluated using 10-fold cross-validation on the dataset proposed by Papadopoulos [98]. and the FVC 2018 dataset with an F1 score of 0.85 and 0.69 respectively. Whereas, the authors of [98], build a classification model using two sets of features(video metadata and comments). Video metadata that specifically considered linguistic features extracted from the video description text and statistics extracted from the video channel. Whereas, the second feature is based on the comments by incorporating a two-level approach. In the first level, features are extracted from the individual comment, later the credibility of each comment is evaluated independently using a pre-trained model proposed by the authors of [106]. These two sets of features are used to train the support vector machine classifier. The algorithm is evaluated using 10 fold cross-validation on their proposed dataset with an F1 score of 0.90 on the fusion of both the features. The other comparative analysis has been done concerning the algorithm proposed by the authors of [107], the method is evaluated on a 70:30 percentage split scheme, where it has been observed that our proposed work outperforms the method given in [107], considering all three measures(Precision, Recall, and F-Measure) by employing FVC dataset. In [107], the author has proposed an algorithm to counter misleading videos as a supervised classification task. A deep learning-based approach UCNet has been developed, along with it some simple features are used for the detection of fake videos. It can be seen that the Decision-Tree, SVM, and Logistic Regression classifiers on the proposed approach outperform the state-of-the-art, except the Random forest classifier. From the above study, it has been observed that most of the reported work mainly employed video metadata and comments based features for the prediction of clickbait videos, however to the best of our knowledge, none of the above-mentioned work has included video related features including video transcript as well as not discussed the similarity among the video title and its content (video transcript), due to which unable to identify how faithfully the video representing the text it claiming to. It has been found from the results outcomes that transcript based features are efficient in improving the model performance and also helps in identifying certain clues about the video credibility that whether it is faithfully representing the same as it claims to. Along with this some of the crucial and novel features are proposed concerning to different feature categories video, comments, and channel that combinedly helps in achieving efficient results. The comparative analysis is shown in Table 1; from Table 1 it can be observed that the proposed algorithm outperforms the existing state-of-the-art methods. The proposed approach outperforms the method proposed by [105][98], concerning both Recall and F-score on 10 fold cross-validation scheme over the FVC dataset. The comparison has been applied to video, comment, and all set of features(fusion), where the proposed model outperforms existing work. On the other hand, while considering the FVC-2018 dataset, it performs better with respect to recall. From the comparative analysis shown in Table 4.2.10, the SVM classifier performs better than the previous approach with respect to recall and f-measure and approximate similar with respect to precision when considering the FVC dataset. In addition to this, the SVM also performs well on our proposed dataset (MVD) with an accuracy of 95.4% on 10-fold cross validation as shown in Table 4.2.10. However, in case of FVC-2018 dataset, it is less effective on video-based and comment based features individually, whereas by applying fusion of features the classifier performs better with respect to recall and f-score compared to previous approaches. The reason for worse performance of the SVM classifier on the FVC-2018 dataset is existence of noise in the feature data. The study also reports that the SVM doesn't perform very well, when the data set has more noise i.e. target classes are overlapping that leads to misclassification of samples. The visualization of feature data is as shown in Fig.4.2.12, where scatter plot matrix is presented as an example for some features to show the distribution of target samples on MVD and FVC-2018 Dataset.

The channel view to subscriber ratio(CS(x)) feature has been visualized w.r.t to four other features (subscriber_to_age_ratio SA(x), video_age_ratio VA(x), fake_comment_count_ratio FCCR(x) and channel views CV(x)) to see the distribution of target sample points. From Fig.4.2.12 it can be noted that the samples of target class are noisy and overlapping in case of FVC-2018 dataset reported in Fig.4.2.12(b) in comparison to MVD dataset reported in Fig.4.2.12(a) and due to which we cannot be able to get a better decision boundary for classification and many of the real samples are misclassified as fake, as a result of which it may not end up performing well.



(a)

(b)

Fig.4.2.12: Plot Matrix representation of features CS(x) against four other features on (a): MVD and (b): FVC-2018 dataset for Clickbait's (Red) and Non-Clickbait's (Green)

Table 4.2.10. Comparative Analysis with the State-of-the-art

Method	Classifier	Split/fold	PRE	REC	FM	Dataset
[105] 2019	SVM(Video feature)	**10 fold**	**0.88**	0.79	0.82	FVC
	SVM(Comment Feature)	**10 fold**	**0.88**	0.74	0.79	FVC
	SVM(Fusion)	**10 fold**	**0.88**	0.82	0.85	FVC
	SVM(Video feature)[YT]	**10 fold**	**0.87**	0.59	**0.70**	FVC-2018
	SVM(Comment Feature)[YT]	**10 fold**	**0.91**	0.53	**0.67**	FVC-2018
	SVM(Fusion)[YT]	**10 fold**	**0.79**	0.61	0.69	FVC-2018
[98] 2017	SVM (Video Feature)	**10 fold**	**0.88**	0.79	0.82	FVC
	SVM (Comment Feature)	**10 fold**	**0.88**	0.74	0.79	FVC
	SVM RBF(Fusion)	**10 fold**	**1.00**	0.83	**0.90**	FVC
[107] 2019	Random Forest	**70:30**	0.74	0.73	0.73	FVC
	Decision- Tree	**70:30**	0.73	0.67	0.67	FVC
	SVM	**70:30**	0.56	0.55	0.54	FVC
	Logistic Regression	**70:30**	0.53	0.53	0.53	FVC
	UCNet	**70:30**	0.82	**0.82**	**0.82**	FVC
Our Method	Random Forest	**70:30**	**0.84**	**0.78**	**0.77**	FVC
	Decision- Tree	**70:30**	**0.77**	**0.75**	**0.73**	FVC
	SVM	**70:30**	**0.65**	**0.63**	**0.63**	FVC
	Logistic Regression	**70:30**	**0.65**	**0.65**	**0.65**	FVC
	SVM (Video Feature)	**10 fold**	0.87	**0.83**	**0.83**	FVC
	SVM (Comment Feature)	**10 fold**	0.87	**0.83**	**0.83**	FVC

	SVM (Fusion)	**10 fold**	0.87	**0.85**	**0.85**	FVC
	SVM(Video Feature)[YT]	**10 fold**	0.57	0.57	0.57	FVC-2018
	SVM(Comment Feature)[YT]	**10 fold**	0.57	**0.57**	0.57	FVC-2018
	SVM (Fusion)[YT]	**10 fold**	0.69	**0.69**	**0.69**	FVC-2018

4.3 Significant Outcome

In this section, we have discussed the proposed frameworks concerning the post accompanied with some multimedia content and here we are particularly more focused on videos and images. Firstly, we have presented a novel and effective method of predicting tweet/ claim accompanying an image to identify how faithfully an image represents a tweet/ claim and to classify them into misleading and real. Using publicly available benchmark verification corpus VMU (2015), we have provided a novel technique via extracting clues from both tweet and image. The five-set of novel clues (Trace of fake concerning to query, Trace of fake concerning to titles, Trace of doubt concerning to query, Trace of doubt concerning to titles, the semantic similarity between title and a query) with respect to tweet and images have been extracted from a tweet and images. The images are processed and effective titles are retrieved. From the study, it has been observed that the retrieval of effective titles plays a major role in improving the performance of the model. The two prominent image search engines are utilized for processing an image (Google Image search and Microsoft Bings visual search). From the comparative analysis, it has been observed that utilizing Microsoft Bings Visual Search is quite more effective in retrieving efficient titles and helps in improving the performance of the model. The results showed that the proposed method outperform the other state-of-the-art methods. In the future, we are more likely to build a solution that can incorporate other multimedia items (Videos, audio, speech attached with tweet/claim) as well as try to build effective real-time application and browser plug-in from a user perspective that can help in the prediction of misleading content in real-time. Secondly, for detecting claims accompanying with some videos, we develop the CVD scheme to detect clickbait videos. The scheme leverages on three components for learning three sets of latent features based on User Profiling, Video-Content, and Human Consensus that further be used to retrieve three sets of cognitive evidence, as an innovative idea for the detection of clickbait videos on YouTube. The set of features are given as an input to the machine learning model and performance has been analyzed by considering all sets of features and each feature independently by employing a different set of the classifier, and it has been observed that J48 outperforms all others with an accuracy of 98.89% by applying all set of features using cross-validation technique, while 97.47% using percentage split technique on the self-generated dataset [MVD]. The proposed method also performs well on the FVC, FVC-2018 dataset, and outperforms the state-of-the-art with an improved Recall and F score. From the analysis, it has been observed that non-clickbait's video channel has more subscribers with respect to registration age as compared to clickbait's video channel. The findings reveal that the average cosine similarity for Non-Clickbait videos is too less than that of the clickbait's, the one of the reasons may be because the fake video probably repeats the same sentence as mentioned in the title, of many times to extend the length of the video with redundant and bogus content.

Further work can be enhanced by generating large datasets and employing more video-related features like image frames from the video to get more efficient clues, as well as the clickbait's headlines, can also be analyzed for other applications like at the time of natural disasters, political elections, healthcare, etc.

Chapter 5

Multi-Web Platform Framework for Fraudulent Content Detection

This chapter presents a novel multi-web platform voting framework that incorporates the 4 sets of novel features (including content features, linguistic features, similarity features, and sentiments features). A unique source platform is designed to collect relevant clues/ headlines from two web platforms (YouTube, Google) based on specific queries and features extracted concerning each collected clue/headline. The effectiveness of the proposed approach is explained and validated through experiments on standard datasets and state-of-the-art comparisons of obtained results.

In this chapter, we discuss the proposed multi-web platform framework for detecting fraudulent/ deceptive claims on social media platforms. Spreading of misleading information on social web platforms has fuelled huge panic and confusion among the public regarding the Corona disease, the detection of which is of paramount importance. Previous studies mainly relied on a specific web platform to collect crucial evidence for the prediction of misleading information. The analysis identifies that retrieving clues from two or more different sources/ web platforms gives more reliable prediction and confidence concerning a specific claim. This study proposed a novel multi-web platform voting framework that incorporates the 4 sets of novel features (including content features, linguistic features, similarity features, and sentiments features). The features have been gathered from each web-platforms to validate the news. To validate the fact/claim, a unique source platform is designed to collect relevant clues/headlines from two web platforms (YouTube, Google) based on specific queries and features extracted concerning each collected clue/headline. This unique platform can also help researchers to gather useful/efficient clues from various web platforms. After evaluation, it has been observed that our proposed intelligent strategy gives promising results and is quite effective in predicting misleading information. The model correctly detected about 98% of the COVID misinformation on the constraint Covid-19 fake news dataset. Furthermore, it is observed that it is efficient to gather clues from multiple web platforms for more reliable predictions to validate the news. The proposed work provides practical implications for the policy-makers and health practitioners that could be useful in protecting the world from misleading information proliferation during this pandemic.

Many people share fake cures to get rid of coronavirus disease without any verification and create lots of misconceptions. Government and officials have also urged people to check the authenticity of the post before sharing [108]. This also motivates us to build an intelligent system for the prediction of fake news spreading during this pandemic. We, therefore, developed a generalized multi-web platform framework of detecting misleading content on social media platforms, where we have considered COVID-19 as a special issue that is a huge pandemic and taken as one of the application case studies in this work. However, our model is generalized and works for other applications as well. COVID-19 is an emerging issue and very few research has been reported yet in this context that leads to motivates us to build an efficient framework to predict misleading content spreading during the COVID outbreak. The major key contributions of the work are highlighted in the following points.

- The proposed work contributes to providing a novel generalized Automated Multi-Web Platform Voting Framework for collecting and validating misleading content in an online social network where considering COVID-19(fake news spreading during Corona outbreak) is one of the special case studies from the application perspective.
- To the best of our knowledge, we are first to build a unique platform (Facts collector) to collect crucial facts and knowledge concerning a claim from two different prominently used social media and web search platforms (YouTube and Google) for validating the claim. Along with this, we provide a different mechanism to search the query (build query) to get efficient and relevant results.
- The four sets of novel features based on content, linguistics/semantic cues, similarity, and sentiments have been extracted/gathered from web platforms that further fed into an ensemble-based machine learning model to classify the news as Misleading or real. In addition, confidence/support has been gathered from different web platforms.
- The COVID-19 is one of the emerging issues. Very few studies have been reported to predict the fake news propagating during this phase, leading to a major contribution by providing the analysis, which greatly helps researchers for further study.
- We evaluate the performance of the model with different classifiers, and a comparative study reveals that the proposed technique outperforms other states of the art approaches on the same dataset.

The primary aim of this chapter is to provide a framework to determine the veracity of any claim posted on social media and to validate whether the given claim is fake or real. To achieve this aim, we propose a multi-web platform framework to gather crucial clues that further be utilized to extract features from the retrieved clues/facts. The four sets of features (including content-based features, linguistic-based features, similarity-based features, and sentiments-based features) are employed for this purpose.

In this work, we investigate the following research questions:

RQ1: Whether incorporating a multi-web platform is effective and more reliable than a single web platform for the prediction of false information?

RQ2: Does the incorporation of effective clues from one platform help improve the model's performance when another platform cannot return the relevant facts?

RQ3: Does every Social web platform is effective in collecting crucial facts concerning a claim?

RQ4: Which one of the features is more effective in discriminating against misleading information.?

This section discusses our automated Multi-Web Platform Voting Framework to Predict Misleading Information Proliferated during the COVID-19 outbreak. Before discussing the proposed architecture model, let's first discuss the problem statement and objective. In this study, we have considered a binary class classification problem. We assume that the posted source claim $c = \{c1, c2, c3 \dots ck\}$ can be divided into two classes $Class = \{M, R\}$: 1) Real (R), namely the posted claim correctly represents the event/situation that it refers to, 2) Misleading (M), namely the posted claim does not correctly represent the event that it refers to. The claims that have been included here are related to coronavirus COVID-19. Take an example of the post related to coronavirus. The post that “chlorine and alcohol products cannot kill viruses within the body” is true information and belongs to the real class since much authoritative and authentic news media have reported relevant news and has also been acknowledged by WHO[38]. While one of the posts says that “coronavirus is caused by 5G technology” which later turned out to be a false rumor, as it lacks genuine support and deviates from scientific principles. So, the aim is to learn a classifier from the labeled feature set, that is $f : X^k \longmapsto Y^k$, where Y^k takes one of the two fine-grained classes: {R, M}. Given the input feature set X^k, the classifier f can output the classification result for the posted claim C^k. We built an automated Multi-Web Platform Voting Framework to Predict Misleading Information Proliferated during the COVID-19 outbreak to address the given problem. The detailed flow diagram of our proposed model is shown in Fig.2. The flow diagram has two phases, Phase 1 and Phase 2. Phase 1, includes the description of data

[38] https://www.indiatoday.in/world/story/drinking-alcohol-will-not-protect-you-from-covid-19-says-who-1653555-2020-03-08

collection and text- pre-processing step, whereas Phase 2, incorporates the detailing of feature extraction and classification. These phases are discussed briefly in the following subsections.

5.1 Data Collection and Text –Pre-processing: Phase 1

The first phase of the proposed architecture incorporates the data collection, query building, facts collection, and text pre-processing step. The data that we have used here for the analysis is collecting samples concerning covid-19 fake news with having binary class labels fake and real. The dataset is the CONSTRAINT 2021 shared task having 5600 real and 5100 fake samples. Each sample from the dataset is considered as an input query that needs to be validated. The input query is passed through the text processing phase where the cleaning of each input sample from the dataset has been performed to make it in a format so that it can be used for further processing, and it includes removal of stopwords, removing duplicates, handling missing data, stemming, punctuation removal, text translation (Google translation API) to the English language, Removing URLs, symbol, emoji, etc. After, text processing the cleaned data is passed to the next module called "Fact Collection" Here, we build a query by adding the " fake news" keyword separated with space. The build query is then passed through a Multi-Web platform to retrieve relevant facts concerning to query. To gather efficient and relevant titles query building is one of the important aspects. What query should be passed to get more relevant responses?. We defined three novel ways to build a query, however from all the given build cases, considered the one most effective in retrieving relevant information, where we have adopted case 1, as others have some limitations as discussed in Table 5.1. Table 5.1, describes 3 possible cases we have considered for building the query. The first case reflects the case of query building after text processing, removing stopwords, and other things, the query is attached with a space concatenated with the "*fake news*" keyword. This case of query generation was found to be good and considered in this study from the analysis. Whereas the other cases include the N_grams concatenated with the keyword "fake news" and the (POS)part of speech tagging(proper nouns), in which we can find all the proper nouns from the input query. Each proper noun is concatenated with the fake news keyword to build a query. However, these cases have certain limitations. Sometimes the context of the query cannot come out properly and miss out on the relevance. The build query is then passed through the two prominent web platforms utilized to retrieve the facts (YouTube platform and Google web search platform). Each collected fact/headline is further going through the text pre-processing part as discussed earlier for cleaning, that later be used to extract/ gather crucial features.

Table 5.1. Possible build cases

Case No.	Possible Cases	Limitation
1.	**Text_preprocessing**(Input_query)+ " "+ "fake news"	-
2.	**N_grams**(Input_query)+ " " + fake news	1. Sometimes the context of the query cannot come out properly and miss out.
3.	**Pos**(Input_query)+" "+ Fake news	1. Sometimes the context of the query cannot come out properly. 2. In some of the cases giving too many irrelevant facts, goes out of context.

Finally, the top 10 title headings are scrapped automatically using selenium from both platforms that further be used for analysis. The algorithm for fact collection is shown below in Algorithm 5.1. Algorithm 5.1, shows the process of facts collected from the Multi-web platform. Here, in this study, we have incorporated two social media and web search platforms for retrieving efficient facts/title heading concerning a query that further is used in feature engineering and validates the claim as fake/real. However, other platforms like Twitter have also been explored for the collection of facts, but the issue with Twitter is it supports keyword-based searching and long query-based search is not applicable, which leads to being a major issue in the collection of relevant facts. Whereas, in the case of google web search and YouTube, we can fetch efficient responses concerning a query.

5.2 Feature extraction and Classification: Phase 2

The second phase is the clues extraction and classification step, and this module takes the facts collected from the previous step and utilizes them to get some efficient clues to predict/classify the claim as fake/real. The four sets of features are employed based on content, linguistics/semantic cues, similarity, and sentiments. Each of these feature categories has been discussed in detail in the below section and in Table 5.2.

5.2.1 Content-based features:

The content-based features have been widely explored in numerous data mining research fields. In this work, we have incorporated content-based features for the prediction of misleading information, including question mark count and fake word count. The question mark count gives necessary clues regarding the confidence reflected in the sentence. If the sentence is showing uncertainty, it means that the claim is not sure regarding that event. Question mark count plays a major role in finding the uncertainty in a given sentence. If any question mark has been encountered in a title/headings retrieved from the web platform while searching a specific query, it returns true. In contrast, fake_word_count is also one of the important features that discriminate fake from real. There is some set of false_phrase_corpus that incorporates a list of keywords that are prominently used to represent fake news.

The keyword corpus that we have created including following phrases: {*‘false','misleading','inaccurate','rumor','rumour', 'not correct' ,'fake news', 'incorrect', 'wrong', 'confounding', 'deceiving', 'deluding', 'wont', 'did' 'Did', 'funny', 'memes', 'catchy', 'bogus', 'counterfeit', 'fabricated', 'fictitious', 'forged', ,'fraudulent', 'mock', 'phony', 'affected', 'artificial', 'erroneous', 'fake', 'fanciful', 'faulty', 'improper', 'invalid', 'mistaken', 'unfounded', 'unreal', 'untrue', 'untruthful', 'casuistic', 'fishy', 'illusive', 'imaginary', 'inexact', 'lying', 'misrepresentative', 'falsity', 'misreport', 'misstatement', 'deception', 'falsification', 'artificial', 'fabrication', 'falsehood', 'hoax', '?', 'Not Died', 'misinformation', 'not committed', 'not dead', 'death rumour', 'is it true', 'not known', 'no proof', 'no known', 'no scientific evidence', 'no evidence', 'not verified', 'clickbait', 'not proven', 'denied', 'deny', 'unverified', 'falsely', 'myth', 'ridiculous', 'not true'*}, if any of these word has been encountered in the retrieved responses corresponding to a query, the fake count incremented by 1. The feature is helpful in identifying fake as the title having these phrases more likely representing news as fake.

Table 5.2: Detailed description of proposed features

Feature Category	Features	Feature Description
Content-Based	• Question mark count • Fake word count	• Number of question mark in a title heading. • Number of fake words encountered in a title heading.
Linguistics/Semantic cues- Based	• NLTK POS TAGGING Semantic Similarity	• The nltk wordnet’s synset is used to measure the semantic similarity between user query and title headings.
Similarity-Based	• Cosine Similarity $\boldsymbol{COS}(\boldsymbol{x},\boldsymbol{y}) = \frac{x.y}{\lvert x\rvert.\lvert y\rvert}$	• The cosine similarity is used to measure the similarity between user query and title headings.
Sentiment-Based	• Query sentiments • Clue sentiments • Sentiment match count	• This features return the sentiment of the user query, either positive negative or neutral. • This features return the sentiment of the title heading, either positive negative or neutral.

		- This features return the count of how many times the query sentiments matchs with the title heading sentiments.

Algorithm 5.1.(Facts Collection)

Input(Text/Text additive image) and Output(Collection of Facts)

```
 def. main ():
  if(image)
      claim= OCR(image)
  else:
  claim= raw input(claim)
  claim= text_preprocessing(claim)
  build query= Stopword_removal(claim) + " "+ “fake news”
  facts google= google_fact_collect (build query)
  facts_youtube= youtube_fact_collect (build query)
 def. google_fact_collect(query):
     for j in search (query, tld="co.in", num=10, stop=10, pause=2):
            f_response= s.get (j, headers = {'User-agent': 'your bot 0.1'})
             soup = BeautifulSoup (f_response. text, 'lxml')
             input_str1 = soup. Find('title'). get_text ()
             input_str1= input_str1.lower()
             input_str1= clean(input_str1)
              return(input_str1)
  def. YouTube_fact_collector(query):
      q= query +" "+'fake news'
      print(q)
      url = "https://www.youtube.com/results?search_query=" + q
      print(url)
      count=0
      driver.get(url)
      user data = driver. find_elements_by_xpath('//*[@id="video-title"]')
      print(user_data)
      links = []
      titles= []
      for i in user data:
        links. append(i.get_attribute('href'))
        titles. append(i.get_attribute('title'))
       for x in titles:
          v_title = x
          return(v_title)
```

5.2.2 Linguistics/ Semantic Cues- based Features

It is challenging to process raw text intelligently as the exact word used in different contexts and order can mean something completely different, however, while using linguistic knowledge can be possible to understand the semantics and the in what context word has been used in a sentence. For a given claim it is very important to understand in what context it is used. The python library nltk.pos_tag is constructed to do the same. When a raw text is passed as an input, it returns an output(doc object) with a variety of annotations. The given document has been parsed and tagged by nltk and there are some statistical model which enables it to predict which tag or label most likely applies in this context also called as POS(part of speech tagging). The concern is to find a particular part of speech, based on both its definition and its context to mark a word in a text using POS. POS tagging also describes the characteristics of lexical terms within a sentence or text that further be used for making predictions/assumptions about semantics. To compute the semantic text similarity between two sentences, we have used POS (Part of speech) text similarity.

There are different POS tags that can be given to each word in a sentence like $(NNS, noun\ plural)(NNP, proper\ noun, singular)(NN, noun, singular)\ etc.$ NLTK POS tagger is employed to assign grammatical information of each word of the sentence. This feature is useful in computing the semantic text similarity between the user query and the clues retrieved from web platforms. The tags generated by nltk.pos_tag are converted to the tag used by wordnet.synsets. The nltk wordnet's synset is used to measure the similarity.

5.2.3 Similarity-based Feature

This is another category of feature used in this work based on similarity. This feature is helpful in segregating relevant titles/heading from all the given responses, as not all responses are useful for validation. To get the efficient performance of the model we need to remove irrelevant titles from the analysis, only those who cross the threshold value are used for analysis. One of the prominently used similarity measures "cosine similarity" has been used in this work to compute the similarity between two sentences irrespective of their size. The sentences are considered as two vectors and the cosine similarity between two vectors is measured in 'θ'. If the angle between two sentences is 0, they are similar, and if θ = 90° they are dissimilar. The formula of calculating the similarity between two sentences x and y can be given as:

$$\boldsymbol{COS}(\boldsymbol{x},\boldsymbol{y}) = \frac{x.y}{|x|.|y|} \tag{5.1}$$

5.2.4 Sentiment based Features:

Sentiment-based features are the fourth set of features employed for the prediction of fake news. Sentiments play an important role in identifying the polarity of the sentence, whether it is showing positive, negative, or neutral sentiments. Here, we have considered 3 features under this category.

1) Query Sentiment: Query Sentiment is a sentiment of the input query given by the user.

2) Title/heading sentiment: This is a sentiment of the responses(title/heading) received as a search result concerning a specific query.
3) Sentiment match counts: From all the 10 responses retrieved from the web platforms, how many times the sentiments of the query and the titles are matched. It also represents whether the sentiment posed by the input query is equivalent to the responses received. It also means that both query and heading are posing the same sentiments and are presented in the same polarity.

All these above-discussed features are briefly shown in Table 5.2 and the Algorithm 5.2 elaborates the complete process of fact validation, where the functions to evaluate the four sets of features are briefly explained that later be fed to ensemble-based classifier for analyzing the performance of the model.

Algorithm 5.2.(Facts Validation)

Input(facts) and Output(status(fake/real))

```
def. main ():
  Linguistic features=  Feature_extraction_lingustic(facts)
  Content_feature  =   Feature_extraction_content(facts)
  Sentiments_features  = Feature_extraction_sentiments(facts)
  Similarity_features =  Feature_extraction_similarity(facts)
  Classification_model_resultl  =  Ensemble_classifier  (Content_features,  Sentiments_features,  Linguistic_features,
Similarity_features)
   Status= classification_model_result
   Print(Status)
def. Feature_extraction_lingustic(facts):
          r. extract_keywords_from_text(input_str1)
           ti=r.get_ranked_phrases ()
           print(ti)
           Y1= listToString(ti)
           s1 = nltk.pos_tag(nltk.word_tokenize(input_str))
           s2 = nltk.pos_tag(nltk.word_tokenize(input_str1))
           Semantic_similarity= similarity (s1, s2)
           return(Semantic_similarity)
 def. Feature_extraction_content(facts):
        r. extract_keywords_from_text(input_str1)
        ti=r.get_ranked_phrases () # To get keyword phrases ranked highest to lowest.
        print(ti)
        Y1= listToString(ti)
        if any (word in Y1 for word in punctuation):
              pun=pun+1
              print ("pun count", pun)
       if any (word in Y1 for word in keyword):
              fake_count=fake_count+1
              print ("The fake count", fake_count)
 def. Feature_extraction_similarity(facts):
           list1 = word tokenize(input_str)
           list2 = word tokenize(input_str1)
           Similarity = cosine_similarity (list1, list2)
           print ("The similarity:", Similarity)
           return(Similarity)
 def. Feature_extraction_sentiments(facts):
             query_sentiment= get_tweet_sentiment(query)
             print("The sentiment of the query",query_sentiment)
             title_sentiment= get_tweet_sentiment(input_str1)
```

```
print ("The sentiment of the title”, title_sentiment)
if (query_sentiment==title_sentiment):
 senti_count=senti_count+1
 print ("The senti_count”, senti_count)
```

5.3 Experimental Analysis and Results

The experimental analysis is performed on publicly available datasets, and different performance measures are adopted (Precision(pre), Recall(rec), F1-score, Accuracy(acc), etc. to measure the effectiveness of the proposed method and lastly presenting the results showing the performance of the proposed model as well as comparative analysis with other States-of-the-art methods. This section covers each of these points in the following subsections.

5.3.1 Constraint-2021 COVID-19 Fake News Detection Dataset: Dataset Description

Here, we have used the constraint-2021 shared task to detect COVID-19 fake news in English[39]. It is a CONSTRAINT-2021 shared task on hostile post-detection, it incorporates two tasks English and Hindi. This dataset is considered in this work for the evaluation of our proposed model. The dataset is collected from various social media like Twitter, Facebook, Instagram, etc. The main objective of this task is to classify a given social media post into Fake/Real. The dataset collects 10,700 manually annotated social media posts and articles of fake and real news on COVID-19 [109]. Some of the examples of fake and real samples concerning to Constraint-2021 fake news dataset is shown in Table 5.3. The dataset is further split into training validation and test sets in the ratio of 3:1:1 as shown in Table 5.4.

[39] https://constraint-shared-task-2021.github.io/

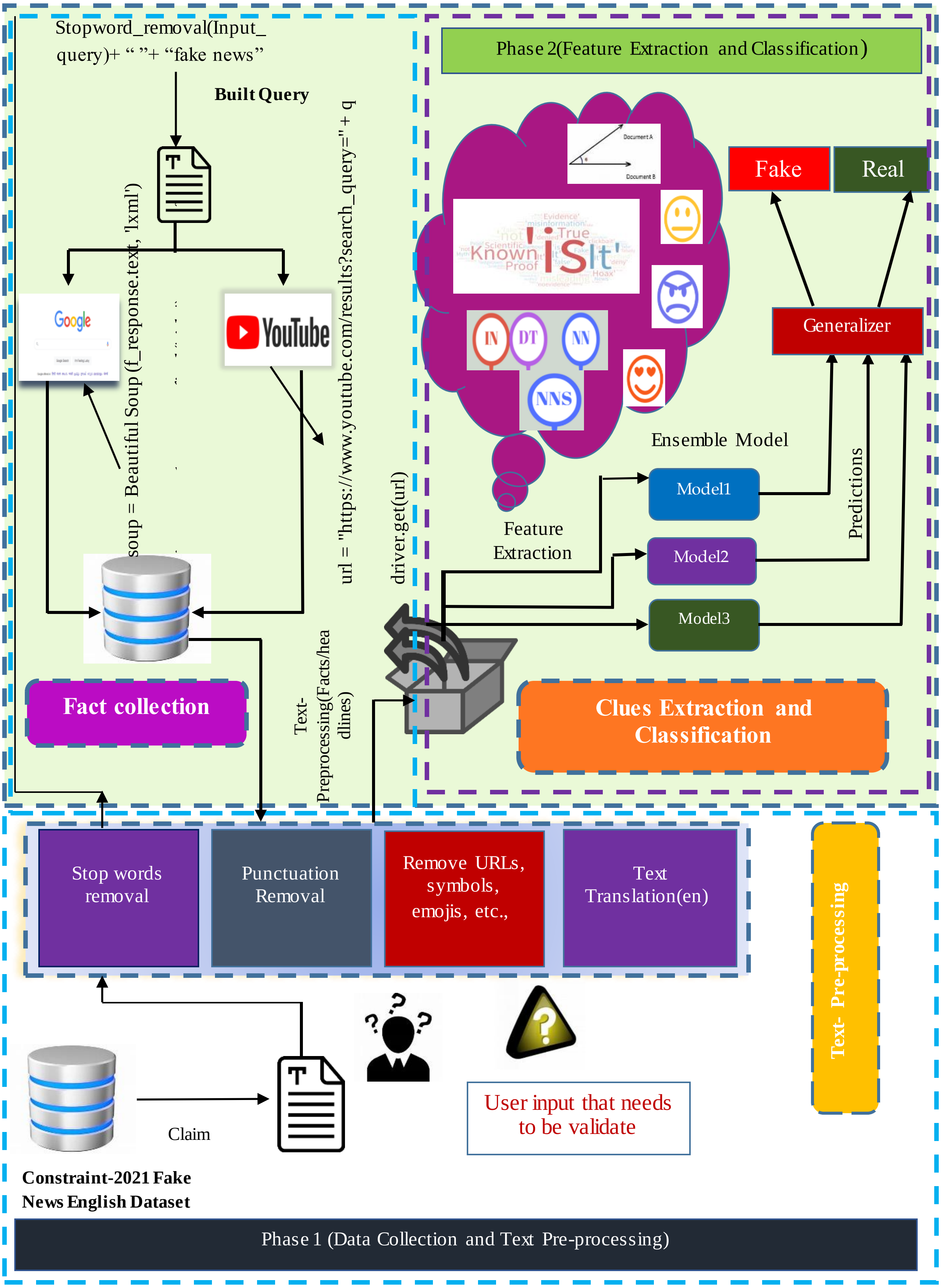


Fig. 5.1. The flow diagram of the proposed approach

Table 5.3. Example of Fake and Real Sample in the Dataset

Label	Text
Fake	Chinese converting to Islam after realising that no muslim was affected by #Coronavirus #COVD19 in the country.
Real	Breathlessness excessive fatigue and muscle aches from COVID can last for months.
Fake	Italian doctor accuse WHO of misleading the world about COVID-19 as COVID-19 is caused by bacteria not virus.
Fake	???Sunlight actually can kill the (novel coronavirus.)??
Real	#COVID19 limits access to the vital support needed for FEP recovery.
Real	The South continues to drive the rising number of COVID-19 deaths. Today 58% of deaths reported were in that region.
Real	You can still fly the friendly skies without fear of COVID if airlines stay serious about safety.

Table 5.4. The Constraint -2021 task dataset description

Split category	Real	Misleading	Total
Training set	3360	3060	6420
Validation set	1120	1020	2140
Test set	1120	1020	2140
Total	5600	5100	10700

5.3.2 Evaluation Measures

To evaluate the performance of the models, we employ four measures of accuracy, precision, recall, and F1-score as our metrics. In addition, these metrics are prominently and widely employ evaluation measures in classification tasks. Each of these measures is explained in Table 5.5.

Table 5.5. Performance Measures

Measure	Definition	Computation Formula
Accuracy	Accuracy can be defined as the proportion of correctly predicted samples to the total number of samples	$\boldsymbol{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$
Precision	Precision is the measure can be calculated as the proportion of truly positive samples in identified positive samples.	$\boldsymbol{Precision} = \frac{TP}{TP + FP}$

Recall	Recall is the performance measure that can be calculated as a proportion of the correctly identified sample in truly positive samples.	$\boldsymbol{recall} = \frac{TP}{TP + FN}$
F1-Score	The F1-score considers the combination of both precision and recall to evaluate the performances.	$\boldsymbol{F1 - score} = \frac{2 * precision * recall}{precision + recall}$

5.3.3 Classification Methods and Results

The Comparative Study with the other state-of-the-art method on the validation set is shown in Table 5.6. The authors of [109] (Model 1) and [110](Model 2), proposed a method to predict misleading information proliferated during the COVID-19 outbreak by employing ensemble-based classification approach, where they reported best run F1-Score of 93.46 by employing SVM and 98.32 using Ensemble-based model respectively as shown in Fig.5.2. and Fig.5.3. respectively. It can be clear that our model outperforms in all discussed cases and provided the best run using ensemble based model incorporating (LR, LSVM and CART) with an F1-Score of 98.88. Whereas, the authors of [111], worked on the same problem task using machine learning, by incorporating various machine learning classifiers and here represented as Model 3. The best run is provided by Model 3 by using SVM with an F1-Score of 95.70, whereas our proposed approach on SVM giving the F1-Score of 98.70 and enhanced the performance by 3% as shown in Fig.5.4.

Table 5.6. Comparative Study with the other state-of-the-art method on the validation set incorporating Google Web Platform.

Method	Model	Precision	Recall	F1-Score	Accuracy
[109] (Model 1)	DT	85.31	85.23	85.25	85.23
[109] (Model 1)	LR	92.76	92.79	92.79	92.75
[109] (Model 1)	SVM	93.46	93.48	93.46	93.46
[110] (Model 2)	Ensemble Model + Heuristic Post-Processing	98.32	98.32	98.32	98.32
[111] (Model3)	SVM	95.71	95.70	95.70	95.70
[111] (Model 3)	LR	95.43	95.42	95.42	95.42
[111] (Model 3)	RF	90.98	90.79	90.80	90.79
[111] (Model 3)	NB	93.33	93.32	93.31	93.32
[111] (Model 3)	MLP	93.62	93.60	93.59	93.60

Our Proposed model	Ensemble voting classifier(LR,CART,LSVM)	98.88	98.88	**98.88**	98.79
Our proposed model	Random Forest	98.20	98.10	98.10	98.09
Our Proposed model	LSVM	98.70	98.70	98.70	98.70
Our Proposed model	Logistic Regression	98.60	98.60	98.60	98.55
Our Proposed Approach model	NB	95.55	95.53	95.54	95.34

Table 5.7. Performance of the model incorporating Google web platform on validation set

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.982	0.981	0.981	0.980
SVM	0.987	0.987	0.987	0.987
LR	0.986	0.986	0.986	0.986
Ensemble learners				
Random Forest	0.986	0.986	0.986	0.985
Voting classifier (RF, LR, KNN)	0.989	0.989	**0.989**	0.989
Voting Classifier (LR, LSVM, CART)	0.989	0.989	0.989	0.987
Bagging Classifier(Decision-Tree)	0.980	0.979	0.979	0.978

Table 5.8. Performance of the model incorporating YouTube web platform on the validation set

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.866	0.852	0.850	0.851
SVM	0.860	0.860	0.860	0.860
LR	0.870	0.869	0.869	0.869
Ensemble learners				
Random Forest	0.866	0.852	0.850	0.851
Voting classifier (RF, LR, KNN)	0.863	0.863	0.863	0.862
Voting Classifier (LR, LSVM, CART)	0.865	0.865	0.865	0.864
Bagging Classifier(Decision-Tree)	0.853	0.795	0.785	0.795

Table 5.9. Performance of the model incorporating both the Web Platform (Google + YouTube) on validation set

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	0.974	0.973	0.973	0.973
SVM	0.980	0.980	0.980	0.979
LR	0.978	0.978	0.978	0.976
SGD	0.980	0.980	0.980	0.975
Ensemble learners				
Random Forest	0.974	0.973	0.973	0.973
Voting classifier (RF, LR, KNN)	0.979	0.979	0.979	0.979
Voting Classifier (LR, LSVM, CART)	0.980	0.980	**0.980**	0.980
Bagging Classifier(Decision-Tree)	0.972	0.971	0.970	0.975

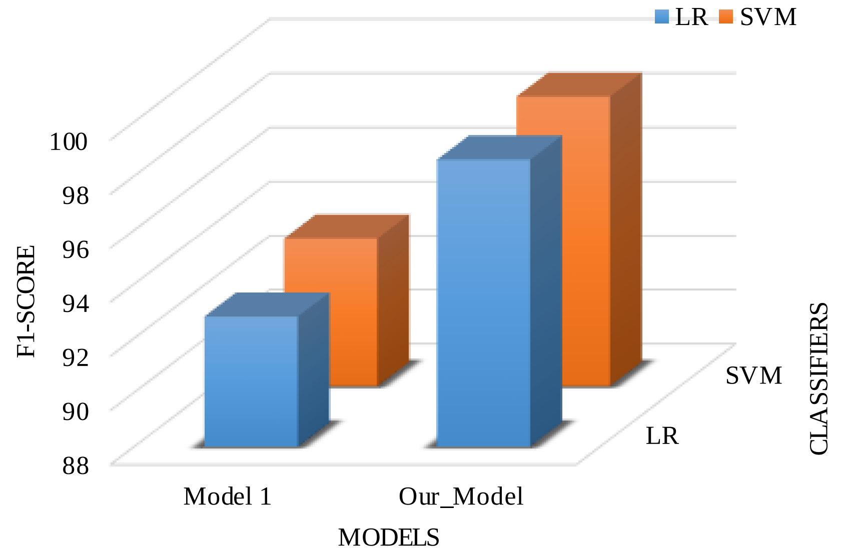


Fig.5.2. Comparative analysis of our with Model 1 on F1 score

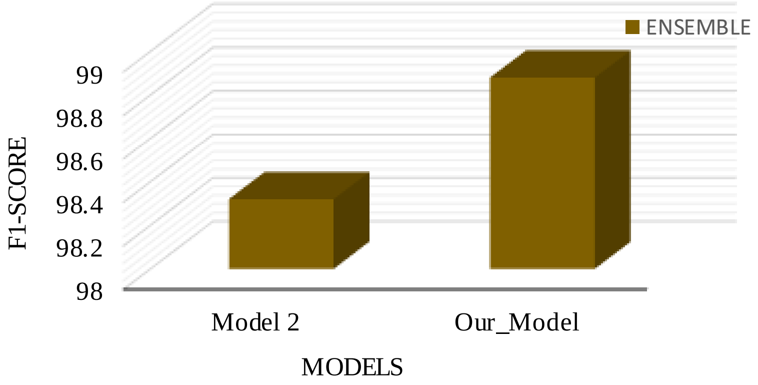


Fig.5.3. Comparative analysis of our model with Model 2 on F1 score

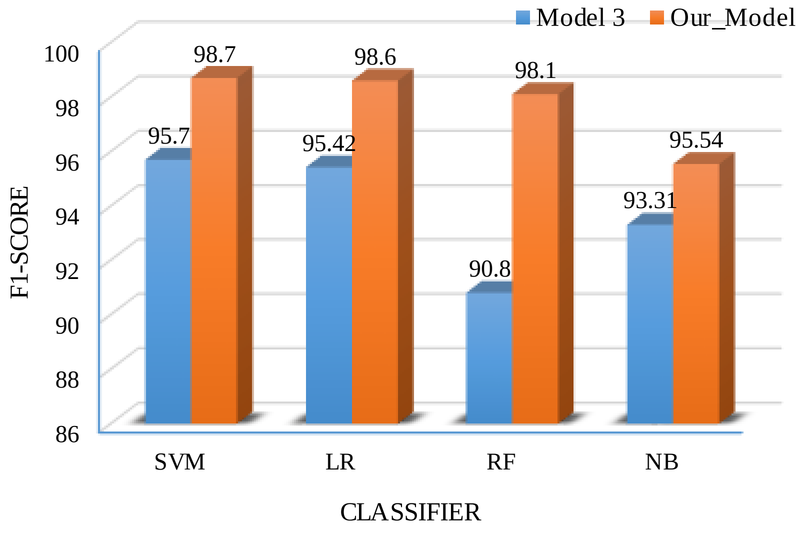


Fig.5.4. Comparative analysis of our model with Model 3 on F1 score

5.3.4 Result Implications and Constraints of the study

The results concerning to this experiment on the validation set are shown in Table 5.7, Table 5.8, and Table 5.9 respectively. From the result analysis, it has been observed that incorporating multi- web platform reduce the uncertainty in many ways. The first case is when one platform is not able to give sufficient clues concerning a claim for the prediction that whether the claim is true or not. Secondly, the supportive clues from different platforms can be one of the reliable sources to validate the claim. The third case, incorporating a multi-web platform improves the performance of the model, like here, we have considered two web platforms YouTube and Google web search. It can be seen from the results discussed earlier that retrieving clues from Google web search independently gives promising results. Whereas YouTube is not performing well independently. So, suppose if we only depend on YouTube to retrieve facts, the model's performance greatly decreases as shown in Table 5.8. To address this case instead of relying on the YouTube platform alone, we have also combined the facts from other platforms (Google Web Search) to validate the facts and improve the model's performance as shown in Table 5.9. From the results, it can also be observed that Google web search was found to be an effective platform for retrieving the crucial information regarding the query. These observations addressed our RQ1, RQ2, and RQ3, and validates that incorporating a multi-web platform is effective and more reliable than a single web platform for the prediction of false information and the incorporation of effective clues from one platform helps improve the performance of the model when another platform is unable to return the relevant facts. As it can be seen, when YouTube alone is not performing well than from the other platform (Google

Search), we can get support to predict the veracity of the claim and improve the model's performance. Moving to RQ4 says that does every social web platform effectively collect crucial facts concerning a claim?. The observation and experimental analysis reveal that it is not true; it is not always mandatory that each platform performs well concerning a query. As some of the platforms may not be able to process the query, there are many other constraints with respect to specific platforms. The real constraint we found in this study is when the web search platform cannot process and understand the query effectively and not give relevant facts concerning to query. In those cases, the prediction is difficult. Here we have considered two platforms for the analysis to address this issue, however, we have also been explored Twitter, but the real constraint we found with Twitter is it only supports keyword-based search, however, the sentence/long query-based search is not applicable, that's why we have not incorporated Twitter in this study.

5.3.5 Feature Evaluation

The four sets of features are evaluated to identify how crucial they are in predicting fake news, the individual performance analysis, and other possible combinations. Table 5.10, describes the possible combination of features and their corresponding results in terms of precision, recall, f1-Score, and accuracy. The proposed four sets of features have been evaluated on the best-run model i.e. Ensemble voting classifier (LR, CART, and LSVM) in our study. It can be seen from Table 5.10, that the (content+similarity) and (Content+linguistic+Sentiment+ Similarity) based features together outperform all other feature combinations in terms of their F1-score with a value of 0.988. Whereas, with respect to the accuracy, the best run is provided by the (content+ similarity) features with a value of 98.83. From these observations, we can now be able to answer RQ4. With respect to research question 4 i.e. Which one of the features is more effective in discriminating against misleading information.?. The (content+ similarity) and (Content+ linguistic+ Sentiment+ Similarity) based features together perform best and are more effective in discriminating misleading information.

Table 5.10. Feature Evaluation

Feature	Accuracy	Precision	Recall	F1-Score	Classifier
Content	98.73	0.987	0.987	0.987	
Linguistic	98.46	0.985	0.985	0.985	
Sentiment	98.51	0.985	0.985	0.985	
Similarity	98.66	0.987	0.987	0.987	

Content+ Linguistic	98.51	0.985	0.985	0.985	Ensemble voting classifier (LR, CART, LSVM)
Content +Sentiment	98.68	0.987	0.987	0.987	
Content+ Similarity	**98.83**	0.988	0.988	**0.988**	
Linguistic + Sentiment	98.5	0.985	0.985	0.985	
Linguistic + Similarity	98.57	0.986	0.986	0.986	
Sentiment+ Similarity	98.51	0.985	0.985	0.985	
Content+ Linguistic+ Sentiment	98.70	0.987	0.987	0.987	
Content + Sentiment+ Similarity	98.70	0.987	0.987	0.987	
Content+linguistic+ Sentiment+ Similarity	98.79	0.988	0.988	**0.988**	

5.3.6 Significant Outcome

This section discussed an intelligent generalized strategy for identifying possible clues to predict misleading information, where fake news proliferated during the COVID-19 outbreak is considered a special case study and detailed analysis has been discussed. We proposed an automated Multi-Web Platform Voting Framework considering YouTube and Google as major sources for the retrieval of clues. The four sets of novel features based on content, linguistics/semantic cues, similarity, and sentiments have been gathered from these platforms that further fed into an ensemble-based machine learning model to classify the news as Misleading or real. Voting is applied to validate the news and to check the confidence/support given by different web platforms. It has been observed that the Google web platform itself performs well in retrieving crucial knowledge, giving the best F1-Score of 98.88 by employing an Ensemble-based model incorporating LR, LSVM, and CART and their voting gives the final decision. However, considering YouTube as a web platform alone for retrieving knowledge it only can give an F1-Score of 86.90 by employing LR which is quite low. Here, we can see YouTube alone is not able to retrieve effective clues to predict the news, however, incorporating a multi web platform scheme we can improve the performance of the model by taking support from other platforms to validate the veracity of news when it is not available. Retrieving clues from a multi-web platform improves the model's performance and outperforms other state-of-the-art techniques on the same dataset by employing an ensemble-based classification model. In the future, one can incorporate and explore other platforms (Instagram, WhatsApp, etc.) to validate the news and expand the work by including different modalities of data (images, videos, etc). Along with this, we are also planning to build a real-time application for the users to predict misleading content.

Chapter 6

Conclusion and Future Scope

In this chapter, we conclude the work, where we briefly discuss each of the previous chapters and the significant findings, along with this we also added the future scope discussing the various other application aspects of this work and explored well that further can be used in further research.

6.1 Conclusion

The thesis covers a diverse set of deceptive activities (Fake news, Hoax, Rumour, Clickbait). Firstly, we discuss the background details of fraudulent content detection, where the understanding concerning to its definition, types, statistics of social media platforms, comparison among various fraudulent activities, motives of doing, motivation, challenges, application, approach, and contribution to the work has been provided. Secondly, we discussed the states of the art techniques provided for the detection of fraudulent content, later on, the research gaps and research objectives have been discussed in detail. In the third chapter, we discussed the proposed techniques build for the detection of fraudulent content, firstly by employing text embedded images and then the text-only part. Whereas, the fourth chapter incorporates text accompanying with some multimedia content (Image or Video). The fifth chapter discusses multi-web platform techniques and how crucial evidence is collected for the detection of fraudulent content. The comparative studies from the other states-of-the-art method reveal that the proposed technique was found to be effective and gives better results.

6.2 Future Work

Many different approaches have been adopted by the researchers to better understand and characterize fraudulent content, this diversification helps to give focus on the future enhancement of the rumor analysis system. Despite substantial advancement in the research field, there are still open research issues that are required for further study. Our findings suggest a need for further research, this consist of the following aspects:

- From most of the studies, it has been observed that the problem of false information detection has been resolved as a binary class (fake or not real) classification problem and the multiclass classification aspect is less explored.

- Very few standard datasets are available in the area of fraudulent content detection, those which are available are mostly from the Twitter web platform, however other platforms like Facebook, YouTube, etc. are less explored. Along with this, there is a dearth of the publically available dataset in the form of images, audio, and video, and needs further attention.

- To lubricate the rumor detection, the unlabelled data should be investigated for unsupervised machine learning models, as labeling of data is labor-intensive.

- Different prominent forms of misleading content are available on social media[112] and used interchangeably concerning different contexts. Across all different categories, it has been observed that hoax and clickbait's are the least addressed area, which requires further attention

- From the study, it has been observed that multimedia data (images, audio, video) plays a major role in the news diffusion, as images, video, or audio are a more convenient and effective way to tell a news report with attached images due to text length limitation. However, very few works have been addressed, Fraudulent content detection problem considering multimedia data or accompanying multimedia content, especially videos and audios. Hence, there is a need for further research in this field.

- From the previous studies, it has been observed that in the case of text accompanying multimedia data, few works have been collected evidence concerning to image forgery.

- The available data on social media is in different languages, there is a need to address the problem of Fraudulent content detection with multilingual content.

These above-discussed aspects extend or improve the current state of knowledge in this field, in a way that, by building the fraudulent detection model on multimodal data, it can be easy for the user to identify the credibility of content in any form, not just with text. As we have seen in the recent example of COVID 19 lots of audio and video messages are propagated to mislead people regarding how to get rid of the coronavirus. One of the audio clips gone viral on WhatsApp and widely shared on Twitter and YouTube, that attributed to Dr. Devi Shetty Chairman and founder of Narayan Health, advises everyone "who has the coronavirus or is suspected of it should not go to get tested[40]", which later turned out to be false. This creates

[40] https://www.altnews.in/fake-audio-clip-attributed-to-dr-devi-shetty-advises-against-getting-tested-for-coronavirus/

lots of misconceptions among the public. The normal user cannot able to verify the news, as there is no real-time tool or extension are available for analyzing multimodal data. However, some tools are available to process text[41]. There are a variety of applications where there is a need to implement an efficient framework for rumor detection, some of the crucial areas are the election, healthcare, natural disaster, terrorism, etc. A recent example of COVID-19 pandemic, where lots of health-related rumors are spreading, pretending to be posted by some government officials[42]. In this situation, where people have an eye over any news announcement related to corona, malicious users use this opportunity, one of the recent news reports that malicious users are sending emails purporting to be from HR departments, executives, and health organizations[43] and playing with human psychology to make them believe that the mail is coming from some government organization. These areas are still open to research.

Another thing is the classification, most of the researchers have tackled the problem as a two-class classification problem, where if some more exploration about the text has been applied it can also be predicted for muti-class, depending upon how much truth the content is presenting and divided it into multiple levels.

[41] http://twitdigest.iiitd.edu.in/TweetCred/
[42] https://www.buzzfeednews.com/article/janelytvynenko/ftc-fda-scam-coronavirus-cures
[43] https://www.buzzfeednews.com/article/janelytvynenko/coronavirus-fake-news-disinformation-rumors-hoaxes

References

[1] M. Alrubaian, M. Al-Qurishi, M. M. Hassan, and A. Alamri, “A Credibility Analysis System for Assessing Information on Twitter,” *IEEE Trans. Dependable Secur. Comput.*, vol. 15, no. 4, pp. 661–674, Jul. 2018.

[2] S. Vosoughi, M. `Neo’ Mohsenvand, and D. Roy, “Rumor Gauge: Predicting the Veracity of Rumors on Twitter,” *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 4, pp. 50:1--50:36, Jul. 2017.

[3] K. Wu, S. Yang, and K. Q. Zhu, “False rumors detection on Sina Weibo by propagation structures,” *2015 IEEE 31st Int. Conf. Data Eng.*, pp. 651–662, 2015.

[4] A. Habib, M. Z. Asghar, A. Khan, A. Habib, and A. Khan, “False information detection in online content and its role in decision making: a systematic literature review,” *Soc. Netw. Anal. Min.*, vol. 9, no. 1, pp. 1–20, 2019.

[5] C. Shen, M. Kasra, W. Pan, G. A. Bassett, Y. Malloch, and J. F. O’Brien, “Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online,” *new media Soc.*, vol. 21, no. 2, pp. 438–463, 2019.

[6] I. Moya, M. Chica, J. L. Sez-Lozano, and scar Cordn, “An Agent-based Model for Understanding the Influence of the 11-M Terrorist Attacks on the 2004 Spanish Elections,” *Know.-Based Syst.*, vol. 123, no. C, pp. 200–216, May 2017.

[7] M. Sun, H.-F. Zhang, H. Kang, G. Zhu, and X. Fu, “Epidemic spreading on adaptively weighted scale-free networks,” *J. Math. Biol.*, vol. 74, 2016.

[8] F. Q. Fu, N. A. Christakis, and J. H. Fowler, “Dueling biological and social contagions,” in *Scientific reports*, 2017.

[9] H. Zhang, M. A. Alim, X. Li, M. T. Thai, and H. T. Nguyen, “Misinformation in Online Social Networks: Detect Them All with a Limited Budget,” *ACM Trans. Inf. Syst.*, vol. 34, pp. 18:1-18:24, 2016.

[10] X. Zhou and R. Zafarani, “A survey of fake news: Fundamental theories, detection methods, and opportunities,” *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–40, 2020.

[11] E. Kochkina, M. Liakata, and A. Zubiaga, “All-in-one: Multi-task Learning for Rumour Verification,” in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3402–3413.

[12] S. Lomborg and A. Bechmann, “Using APIs for Data Collection on Social Media,” *Inf. Soc.*, vol. 30, no. 4, pp. 256–265, 2014.

[13] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, “Detection and Resolution of Rumours in Social Media: A Survey,” *ACM Comput. Surv.*, vol. 51, no. 2, pp. 32:1--32:36, Feb. 2018.

[14] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, “Rumor Has It: Identifying Misinformation in Microblogs,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1589–1599.

[15] Y. Liu and S. Xu, “Detecting Rumors Through Modeling Information Propagation

Networks in a Social Media Environment," *IEEE Trans. Comput. Soc. Syst.*, vol. 3, no. 2, pp. 46–62, Jun. 2016.

[16] K. Driscoll and S. Walker, "Working within a black box: Transparency in the collection and production of big twitter data," *Int. J. Commun.*, vol. 8, no. 1, pp. 1745–1764, 2014.

[17] S. Santhoshkumar and L. D. D. Babu, "Earlier detection of rumors in online social networks using certainty-factor-based convolutional neural networks," *Soc. Netw. Anal. Min.*, vol. 10, no. 1, pp. 1–17, 2020.

[18] C. M. M. Kotteti, X. Dong, and L. Qian, "Ensemble Deep Learning on Time-Series Representation of Tweets for Rumor Detection in Social Media," *arXiv Prepr. arXiv2004.12500*, 2020.

[19] L. M. S. Khoo, H. L. Chieu, Z. Qian, and J. Jiang, "Interpretable Rumor Detection in Microblogs by Attending to User Interactions," *arXiv Prepr. arXiv2001.10667*, 2020.

[20] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun, "CED: Credible early detection of social media rumors," *IEEE Trans. Knowl. Data Eng.*, 2019.

[21] M. Z. Asghar, A. Habib, A. Habib, A. Khan, R. Ali, and A. Khattak, "Exploring deep neural networks for rumor detection," *J. Ambient Intell. Humaniz. Comput.*, pp. 1–19, 2019.

[22] S. M. Alzanin and A. M. Azmi, "Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation--maximization," *Knowledge-Based Syst.*, p. 104945, 2019.

[23] A. E. Fard, M. Mohammadi, Y. Chen, and B. de Walle, "Computational Rumor Detection Without Non-Rumor: A One-Class Classification Approach," *IEEE Trans. Comput. Soc. Syst.*, vol. 6, no. 5, pp. 830–846, 2019.

[24] A. Zubiaga, M. Liakata, and R. Procter, "Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media," 2016.

[25] S. Kwon, M. Cha, and K. Jung, "Rumor Detection over Varying Time Windows," *PLoS One*, vol. 12, pp. 1–19, 2017.

[26] Y. Liu *et al.*, "Supervised Group Embedding for Rumor Detection in Social Media," in *Web Engineering*, 2019, pp. 139–153.

[27] A. Kumar, V. Singh, T. Ali, S. Pal, and J. Singh, "Empirical Evaluation of Shallow and Deep Classifiers for Rumor Detection," in *Advances in Computing and Intelligent Systems*, Springer, 2020, pp. 239–252.

[28] S. Han, J. Gao, and F. Ciravegna, "Data Augmentation for Rumor Detection Using Context-Sensitive Neural Language Model With Large-Scale Credibility Corpus," 2019.

[29] X. Lin, X. Liao, T. Xu, W. Pian, and K.-F. Wong, "Rumor Detection with Hierarchical Recurrent Convolutional Neural Network," in *CCF International Conference on Natural Language Processing and Chinese Computing*, 2019, pp. 338–348.

[30] J. Ma *et al.*, "Detecting Rumors from Microblogs with Recurrent Neural Networks," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 3818–3824.

[31] S. Hamidian and M. T. Diab, "Rumor Detection and Classification for Twitter Data,"

ArXiv, vol. abs/1912.0, 2019.

[32] S. Han, J. Gao, and F. Ciravegna, "Neural Language Model Based Training Data Augmentation for Weakly Supervised Early Rumor Detection." 2019.

[33] A. Dang, A. Moh'd, A. Islam, and E. Milios, "Early Detection of Rumor Veracity in Social Media," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[34] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee, "Unsupervised Rumor Detection Based on Users Behaviors Using Neural Networks," *Pattern Recogn. Lett.*, vol. 105, no. C, pp. 226–233, Apr. 2018.

[35] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor Detection with Hierarchical Social Attention Network," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 943–951.

[36] L. Poddar, W. Hsu, M. Lee, and S. Subramaniyam, "Predicting Stances in Twitter Conversations for Detecting Veracity of Rumors: A Neural Approach," 2018, pp. 65–72.

[37] J. Ma, W. Gao, and K.-F. Wong, "Rumor Detection on Twitter with Tree-structured Recursive Neural Networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1980–1989.

[38] J. Ma, W. Gao, and K.-F. Wong, "Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 708–717.

[39] C. T. Duong, Q. V. H. Nguyen, S. Wang, and B. Stantic, "Provenance-Based Rumor Detection," in *Databases Theory and Applications*, 2017, pp. 125–137.

[40] T. Nguyen, "A Comprehensive Low and High-level Feature Analysis for Early Rumor Detection on Twitter," *CoRR*, vol. abs/1711.0, 2017.

[41] A. Y. M. Floos, "Arabic Rumours Identification By Measuring The Credibility Of Arabic Tweet Content," *Int. J. Knowl. Soc. Res.*, vol. 7, no. 2, pp. 72–83, Apr. 2016.

[42] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1395–1405.

[43] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: {A} Real-time Web-based System for Assessing Credibility of Content on Twitter," *CoRR*, vol. abs/1405.5, 2014.

[44] S. Mohd Shariff, X. Zhang, and M. Sanderson, "User Perception of Information Credibility of News on Twitter," in *Advances in Information Retrieval*, 2014, pp. 513–518.

[45] G. Cai, H. Wu, and R. Lv, "Rumors detection in Chinese via crowd responses," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 2014, pp. 912–917.

[46] X. Xia, X. Yang, C. Wu, S. Li, and L. Bao, "Information Credibility on Twitter in Emergency Situation," in *Intelligence and Security Informatics*, 2012, pp. 45–59.

[47] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic Detection of Rumor on Sina Weibo," in *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 2012, pp. 13:1--13:7.

[48] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684.

[49] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent Features of Rumor Propagation in Online Social Media," *2013 IEEE 13th Int. Conf. Data Min.*, pp. 1103–1108, 2013.

[50] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman, "Challenges of Computational Verification in Social Multimedia," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 743–748.

[51] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, "SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours," *CoRR*, vol. abs/1704.0, 2017.

[52] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 795–816.

[53] A. Olteanu, S. Vieweg, and C. Castillo, "What to Expect When the Unexpected Happens: Social Media Communications Across Crises," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, 2015, pp. 994–1009.

[54] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," *Proc. 4th ACM Int. Conf. Web Search Data Mining, WSDM 2011*, pp. 177–186, 2011.

[55] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel Visual and Statistical Image Features for Microblogs News Verification," *Trans. Multi.*, vol. 19, no. 3, pp. 598–608, Mar. 2017.

[56] E. Ardizzone, A. Bruno, and G. Mazzola, "Copy–Move Forgery Detection by Matching Triangles of Keypoints," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 10, pp. 2084–2094, Oct. 2015.

[57] W. Wang, "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection," 2017.

[58] D. Varshney and D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," *Appl. Intell.*, pp. 1–22.

[59] Z. Yang, C. Wang, F. Zhang, Y. Zhang, and H. Zhang, "Emerging Rumor Identification for Social Media with Hot Topic Detection," in *2015 12th Web Information System and Application Conference (WISA)*, 2015, pp. 53–58.

[60] T. Takahashi and N. Igata, "Rumor detection on twitter," 2012, pp. 452–457.

[61] M. Mathioudakis and N. Koudas, "TwitterMonitor: trend detection over the twitter

stream," in *SIGMOD Conference,* 2010.

[62] Q. Zhang, S. Zhang, J. Dong, J. Xiong, and X. Cheng, "Automatic Detection of Rumor on Social Network," in *Natural Language Processing and Chinese Computing*, 2015, pp. 113–122.

[63] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the Role of Visual Content in Fake News Detection," *arXiv Prepr. arXiv2003.05096*, 2020.

[64] Y. Geng, J. Sui, and Q. Zhu, "Rumor Detection of Sina Weibo Based on SDSMOTE and Feature Selection," in *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA),* 2019, pp. 120–125.

[65] S. A. Alkhodair, S. H. H. Ding, B. C. M. Fung, and J. Liu, "Detecting breaking news rumors of emerging topics in social media," *Inf. Process. Manag.*, 2019.

[66] A. R. Setlur, "Semi-Supervised Confidence Network aided Gated Attention based Recurrent Neural Network for Clickbait Detection," *arXiv Prepr. arXiv1811.01355,* 2018.

[67] A. Geçkil, A. A. Müngen, E. Gündogan, and M. Kaya, "A clickbait detection method on news sites," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* 2018, pp. 932–937.

[68] K. Shu, S. Wang, T. Le, D. Lee, and H. Liu, "Deep headline generation for clickbait detection," in *2018 IEEE International Conference on Data Mining (ICDM),* 2018, pp. 467–476.

[69] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," *arXiv Prepr. arXiv1704.07506*, 2017.

[70] A. Ishak, Y. Y. Chen, and S.-P. Yong, "Distance-based hoax detection system," in *2012 International Conference on Computer \& Information Science (ICCIS),* 2012, vol. 1, pp. 215–220.

[71] S. Kumar and K. M. Carley, "Tree LSTMs with Convolution Units to Predict Stance and Rumor Veracity in Social Media Conversations," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 5047–5058.

[72] G. Giasemidis *et al.*, "Determining the Veracity of Rumours on Twitter," *CoRR*, vol. abs/1611.0, 2016.

[73] J. Sampson, F. Morstatter, L. Wu, and H. Liu, "Leveraging the Implicit Structure Within Social Media for Emergent Rumor Detection," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 2377–2382.

[74] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect Rumors Using Time Series of Social Context Information on Microblogging Websites," 2015.

[75] S. Hamidian and M. Diab, "Rumor Detection and Classification for Twitter Data," 2015.

[76] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 729–736.

[77] D. K. Vishwakarma, D. Varshney, and A. Yadav, "Detection and veracity analysis of fake news via scrapping and authenticating the web search," *Cogn. Syst. Res.*, vol. 58, pp. 217–229, 2019.

[78] N.-M. Chidiac, P. Damien, and C. Yaacoub, "A robust algorithm for text extraction from images," in *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, 2016, pp. 493–497.

[79] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris, "Detection and visualization of misleading content on Twitter," *Int. J. Multimed. Inf. Retr.*, vol. 7, no. 1, pp. 71–86, 2018.

[80] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manag.*, 2019.

[81] K. Shu, S. Wang, and H. Liu, "Exploiting tri-relationship for fake news detection," *arXiv Prepr. arXiv1712.07709*, vol. 8, 2017.

[82] B. D. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[83] C. Zhu and G. Wu, "Research and analysis of search engine optimization factors based on reverse engineeing," in *2011 Third International Conference on Multimedia Information Networking and Security*, 2011, pp. 225–228.

[84] D. Vishwakarma, D. Varshney, and A. Yadav, "Detection and Veracity analysis of Fake News via Scrapping and Authenticating the Web Search," *Cogn. Syst. Res.*, vol. 58, 2019.

[85] J. Y. Khan, M. Khondaker, T. Islam, A. Iqbal, and S. Afroz, "A Benchmark Study on Machine Learning Methods for Fake News Detection," *arXiv Prepr. arXiv1905.04749*, 2019.

[86] V. L. Rubin and T. Lukoianova, "Truth and deception at the rhetorical structure level," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 5, pp. 905–917, 2015.

[87] C. Janze and M. Risius, "Automatic Detection of Fake News on Social Media Platforms.," in *PACIS*, 2017, p. 261.

[88] A. Bondielli and F. Marcelloni, "A Survey on Fake News and Rumour Detection Techniques," *Inf. Sci. (Ny).*, vol. 497, 2019.

[89] C. Boididou *et al.*, "Verifying Multimedia Use at MediaEval 2015.," in *MediaEval*, 2015.

[90] S. Middleton, "Extracting attributed verification and debunking reports from social media: mediaeval-2015 trust and credibility analysis of image and video," 2015.

[91] Z. Jin, J. Cao, Y. Zhang, and Y. Zhang, "MCG-ICT at MediaEval 2015: Verifying Multimedia Use with a Two-Level Classification Model.," in *MediaEval*, 2015.

[92] C. Boididou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, and Y. Kompatsiaris, "The CERTH-UNITN Participation@ Verifying Multimedia Use 2015.," in *MediaEval*, 2015.

[93] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and I.

Kompatsiaris, "Detection and visualization of misleading content on Twitter," *Int. J. Multimed. Inf. Retr.*, 2017.

[94] D. S. Sisodia, "Ensemble Learning Approach for Clickbait Detection Using Article Headline Features," *Informing Sci. Int. J. an Emerg. Transdiscipl.*, vol. 22, pp. 31–44, 2019.

[95] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, Y. Kompatsiaris, and D. Teyssou, "InVID Fake Video Corpus v2.0." Zenodo, Jan-2018.

[96] D. Y. Zhang, L. Song, Q. Li, Y. Zhang, and D. Wang, "Streamguard: A bayesian network approach to copyright infringement detection problem in large-scale live video sharing systems," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 901–910.

[97] S. A. Papadopoulos, "Towards Automatic Detection of Misinformation in Social Media."

[98] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, and Y. Kompatsiaris, "Web video verification using contextual cues," in *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*, 2017, pp. 6–10.

[99] M. Potthast, S. Köpsel, B. Stein, and M. Hagen, "Clickbait detection," in *European Conference on Information Retrieval*, 2016, pp. 810–817.

[100] L. Shang, D. Y. Zhang, M. Wang, S. Lai, and D. Wang, "Towards reliable online clickbait video detection: A content-agnostic approach," *Knowledge-Based Syst.*, vol. 182, p. 104851, 2019.

[101] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, 2017.

[102] K. Shu, S. Wang, and H. Liu, "Understanding user profiles on social media for fake news detection," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018, pp. 430–435.

[103] K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu, "The role of user profiles for fake news detection," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 436–439.

[104] K. Shu *et al.*, "Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2018, vol. 8, pp. 430–435.

[105] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, and I. Kompatsiaris, "A corpus of debunked and verified user-generated videos," *Online Inf. Rev.*, 2019.

[106] C. Boididou, S. Papadopoulos, M. Zampoglou, L. Apostolidis, O. Papadopoulou, and Y. Kompatsiaris, "Detection and visualization of misleading content on Twitter," *Int. J. Multimed. Inf. Retr.*, vol. 7, no. 1, pp. 71–86, 2018.

[107] P. Palod, A. Patwari, S. Bahety, S. Bagchi, and P. Goyal, "Misleading Metadata Detection on YouTube," in *European Conference on Information Retrieval*, 2019, pp. 140–147.

[108] W. Zhou, A. Wang, F. Xia, Y. Xiao, and S. Tang, "Effects of media reporting on mitigating spread of COVID-19 in the early phase of the outbreak," 2020.

[109] P. Patwa *et al.*, "Fighting an infodemic: Covid-19 fake news dataset," *arXiv Prepr. arXiv2011.03327,* 2020.

[110] S. D. Das, A. Basak, and S. Dutta, "A Heuristic-driven Ensemble Framework for COVID-19 Fake News Detection," *arXiv Prepr. arXiv2101.03545,* 2021.

[111] T. Felber *et al.*, "Constraint 2021: Machine Learning Models for COVID-19 Fake News Detection Shared Task," *arXiv Prepr. arXiv2101.03545,* 2021.

[112] S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis, "The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans," *J. Data Inf. Qual.,* vol. 11, no. 3, pp. 1–37, 2019.