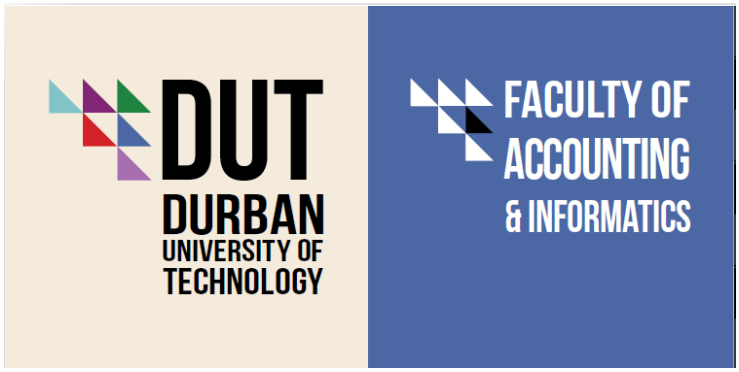


DURBAN UNIVERSITY OF TECHNOLOGY

An Advanced Ensemble Approach for Detecting Fake News

By

Arvin Hansrajh

22065138

A dissertation submitted in fulfilment of the requirement for the Masters in Information and Communications Technology degree

Faculty of Accounting and Informatics, Department of Information Technology, Postgraduate Studies

Supervisor: Dr T. T. Adeliyi

Co-Supervisor: Dr J. W. Wing

2021

DECLARATION

I, *Arvin Hansrajh*, declare that:

(i) The research reported in this dissertation, except where otherwise indicated, is my original research.

(ii) This dissertation has not been submitted for any degree or examination at any other university.

(iii) This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

- Their words have been re-written but the general information attributed to them has been referenced.
- Where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the Reference Section of this disertation.

6 August 2021

Arvin Hansrajh Date

Approved for Final Submission

Supervisor: _ 11/12/2021

Dr T. T. Adeliyi Date

Co-Supervisor 12/12/2021

Dr J. W. Wing Date

DEDICATION

To my late parents, **Mr Hansrajh Sewpaul and Mrs Mayapathee Sewpaul**, who inculcated in me the significance of integrity, modesty, curiosity, resilience, gratitude, hard work and lifelong learning.

To my wife, **Shalina**, for igniting my interest in fake news and continuously stoking my enthusiasm during the entire research journey.

To my only child, my dearest son, **Aarav.**
May this study inspire you to become the best version of yourself.

ACKNOWLEDGEMENTS

I am grateful to several people for their guidance and support during the research and writing of this dissertation. In particular, I would like to thank the following persons.

- Dr T T Adeliyi, my supervisor for his professional guidance. His expertise, insights, support and critical comments were invaluable.

- Dr J W Wing, my co-supervisor, for her advice, encouragement and optimism throughout this study.

- To my family for their love, inspiration, and giving me the inner strength to persevere and complete this project.

Finally, an undertaking of this nature is not possible without the grace of the Almighty. I am indebted to the Creator of the Universe for granting me the strength, courage, and sagacity to accomplish this study.

Arvin Hansrajh
Durban
2021

ABSTRACT

The explosive growth in fake news has evolved into a major threat to society, public trust, democracy and justice. The easy dissemination and sharing of information online provides the unabated momentum. As such, it has become crucial to combat the menace of fake news and to mitigate its consequences. Detecting fake news is an intricate problem since it can appear in a multitude of forms, thus making it both automatically and manually very challenging to successfully recognise. Furthermore, fake news is intentionally created to mislead and is often interspersed with real news.

Studies have shown that human beings are somewhat unsuccessful in identifying deception. The majority of people accept that information they are presented with in virtually any form is reliable or veracious. The relevant literature reveals that a considerable number of people who read fake news stories report that they find them more believable than news that is disseminated via mainstream media. F urthermore, there are predictions that by 2022, the greater population within mature economies are likely to consume more false than true information. The importance of combatting fake news has been starkly demonstrated during the current Covid-19 crisis. Social media networks are significantly increasing their efforts to develop fake news detection mechanisms, as well as to enlighten subscribers on how to recognise fake news, however most people are naturally predisposed to spreading sensationalist news without any fact-checking process in place. It is therefore evident that the creation of automated solutions is vital and urgent for the detection of untruthful news and as such, the goal of this study is to aid in the detection of fake news. Prior studies have included many machine learning models with varying degrees of success but many non-conventional machine learning models have not yet been exploited despite evidence to suggest that they are the best in several text classification scenarios. Consequently, an ensemble learning approach is suggested to assist in resolving the gap that has been identified.

Contemporary studies are validating the efficiency of ensemble learning methods and have provided encouraging outcomes. This study investigates how machine learning and natural language processing methods are pooled together in a blended ensemble in order to build a model that will utilise data from past news articles, to forecast whether a current news article is likely to be false or true. A variety of performance metrics such as roc, roc auc, recall, precision, f1-score and accuracy are used in comparing the proposed model to other machine learning models. The measurements are applied in evaluating and gauging the efficiency of the proposed model. The results obtained show that the proposed model's performance is better than several other learning models, which is very encouraging.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION AND BACKGROUND

1.1 Introduction

The widespread use of social media and the internet has allowed more individuals to acquire news more conveniently from a broader spectrum of online sources rather than just conventional news media that was previously relied upon. This has now become the 'new normal'. Generally speaking, those who are frequently on the internet are more inclined to obtain current event updates and news via social media, thereby increasing their risk of being exposed to wide-scale misinformation (Saeed *et al*., 2020). The primary challenge is that bogus news can be fashioned and produced cheaper, and faster online in comparison to conventional news media (Shu *et al*., 2017). This presents a fecund ground for deception and fake news in the shape of news articles, reviews, hoaxes, advertisements, rumours, exaggerated claims, and satires to flourish. The proliferation of fabricated news has the capacity to blosson into highly harmful consequences for both individuals and humanity (Zhou *et al*., 2019). Evidence suggests since becoming part of our daily lives, that fake news has increased prejudice among people of dissimilar cultural and ethnic contexts, fuelled political violence and negatively influenced issues related to public health, such as the Covid-19 vaccine rollouts. The motivation behind fake news is often to promote certain philosophies and ideas that are usually related to political agendas. Consequently, governments globally are striving to monitor and tackle this challenging problem (Roy *et al*., 2018).

Bogus news is not a recent phenomenon or a new concept (de Beer & Matthee, 2020), nor is it a unique creation of the digital communication era (Rubin *et al*., 2015), however its prolificacy and power has grown exponentially in the last decade. It came to the fore during the US Presidential election of 2016 with the fabrication of several deceptive stories that had political consequences both for that country and the world at large. The majority of media houses positioned Facebook at the centre of the blame for this controversy as Facebook allows uncensored views of the populace to be widely and freely shared (Shukla *et al*., 2019). Facebook's defence is that it was created as a free platform for any individual with access to the internet to be able to share their personal views and opinions. Nonetheless, there is now a general consensus that misinformation and fake news, in particular, has become a huge problem, which could result in significant social costs in the future (Allcott & Gentzkow, 2017). There are now various documented cases where ingeniously

compiled fake news has had grave consequences by inciting ethnic or religious groups into violence (Khan *et al*., 2019). More recently, fake news touted the effectiveness of chloroquine in the treatment of Covid-19, which led to increasing cases of chloroquine overdoses (Busari & Adebayo, 2020).

The omnipresent character of the internet permits anyone to propogate biased and false information effortlessly. In addition, it is practically impossible to prohibit or regulate the production and circulation of fake news. As a result, researchers and online platforms are taking the initiative to proactively deal with potentially fake news. This is a complicated and difficult problem as fake news appears in a multitude of forms, thus making it both automatically and manually very challenging to successfully recognise (Bondielli & Marcelloni, 2019). Headlines as 'clickbait' are regularly used to lure users to view often subjective news articles so as to generate revenue. According to Wang (2017, p. 1), "The problem of fake news detection is more challenging than detecting deceptive reviews, since the political language on TV interviews, posts on Facebook and Twitters are mostly short statements". Furthermore, the global problem of deceptive news is particularly demanding to combat in the existing digital world since fake news or misinformation can propagate easily through a large number of platforms available for sharing information (Agarwal *et al*., 2019).

To a large extent, human beings are rather unsuccessful in identifying deception. Many people have confidence in most information that is presented to them in any form and accept it as reliable or veracious, thereby exhibiting a general gullibility (Pennycook *et al*., 2015) that is marked by a very keen receptiveness to ideas that they do not necessarily even understand. In addition, confirmation bias may influence people to perceive strictly what they prefer. According to Ahmed *et al*. (2017, p. 11), a "sizable number of people who read fake news stories have reported that they believe them more than news from mainstream media". Gartner (2017) has even predicted that "by 2022, most people in mature economies will consume more false information than true information".

Few people have direct access to the verifiable facts related to an event, and so most will simply trust the news forwarded to them by relatives or associates (Sangamnerkar *et al*., 2020). The issue

of ascertaining what are actually verifiable facts within a landscape of so many juxtaposed political and economic agendas is thus really at the root of the fake news problem but addressing this is beyond the scope of this study. Thus, it is apparent that the creation of automated solutions to detect fake news is vital and urgent (Roy *et al*., 2018). Past works have included several classical learning models. Numerous non-conventional machine learning models however have not been employed despite having excelled in several classification of text scenarios (Khan *et al*. 2019). Consequently, an ensemble learning approach is suggested to assist in resolving the gap identified.

Contemporary studies are validating the efficacy of ensemble learning with promising prospects (Gutierrez-Espinoza *et al*. 2020). This particular study investigates how machine learning and natural language processing methods can be pooled together in a blended ensemble in order to build a model that will utilise data from past news articles, to forecast whether a current news article is likely to be false or true. A variety of performance metrics such as roc, roc auc, recall, precision, f1-score and accuracy will be used to evaluate the effectiveness of the proposed model against other machine learning models.

1.2 Research Problem Statement

In recent years, ‘fake news’ has become a colloquially accepted ‘buzzword’ but a universal definition of ‘fake news’ does not currently exist (Khan *et al*., 2019). According to Zhou *et al*. (2019, p. 4), fake news is defined as “intentionally false news published by a news outlet”. Furthermore, from a broader perspective, these authors state that “fake news is false news”. Allcott & Gentzkow (2017, p. 4) contend that “fake news is a news article that is intentionally and verifiably false”. This definition of fake news is adopted in the study. More specifically, a fake news article is deliberately created to misinform or deceive the reader and can be confirmed as bogus by additional sources (Bondielli & Marcelloni, 2019).

As has been mentioned, humans are not good at all at identifying bogus news as evidenced by a success rate of just 54% in the detection of false news, and just 4% if they hazard a guess (Okoro *et al*., 2018). What makes identification of fake news more difficult is that it does not always take the form of a traditional news article. The importance of combatting false news is being clearly illustrated during the current pandemic (de Beer & Matthee 2020). The problem starts when people

in large numbers begin to believe fake news without verifying its truthfulness. Consequently, it is spread further and given more momentum by a large number of social messaging sites and sharing applications (Shukla *et al*., 2019). False news is considered a great threat to democracy, freedom of the press, and the freedom to express oneself individually (Zhou *et al*., 2019).

Due to the prolificacy of 'news', people are often unable to cross-check references, validate and ensure the credibility of said news. This process is time-consuming and involves skills and expertise not usually possessd by the average person. There are currently very few tools readily available to enable the public to adequately test the authenticity of news (Khan *et al*. 2019). As such, the problems caused by fake news continue to persist despite numerous efforts and research studies. Therefore, reliable automated fake news detection is vital. As previously mentioned, classical machine learning has been applied in detecting misinformation with varying degrees of success but many advanced models have not been applied. Current studies are highlighting the potential of ensemble learning (Gutierrez-Espinoza *et a*l., 2020) and as such this study adopts an ensemble strategy to address the gap.

1.3 Research Question

The efficient detection of fake news continues to be a challenging problem to solve by employing both manual and automated methods. This study seeks to provide an appropriate solution to the following research question based on the identified research problem discussed in the previous section of this dissertation. In summary, considering the aforementioned challenges encountered in identifying and detecting fake news, the researcher wishes to address the following primary research question:

What ensemble approach could be developed to ameliorate the difficulties encountered in the reliable detection of fake news?

1.4 Aim and Objectives

The aim of this research is to detect fake news by utilizing an ensemble approach to combat the proliferation and propagation of fake news.

To accomplish this particular aim, the research objectives are set as follows:

1. To comprehensively research relevant publications based on the detection of fake news in order to identify effective methods and approaches to devising a probable solution.

2. To develop an ensemble approach to detect fake news by combining machine learning with natural language processing.

3. To experimentally evaluate the developed model against other machine learning algorithms using well known statistical evaluation metrics.

1.5 Research Methodology

The methodological steps to be followed towards the realisation of the set aim and objectives of this study comprise three sequential stages. The initial stage addresses the first objective of this study by conducting a comprehensive literature review with a focus on identifying methods and approaches to devising a probable solution to the problem. The second stage addresses objective two of this study by elaborating on design research that is relevant to the aim of this study. The proposed framework is presented with a detailed account of the various processes involved at each level. The fake news examples used as case studies were sourced from two publicly available benchmark fake news datasets. GloVE word embeddings and n-grams were selected for feature extraction. Cross-validation was chosen as a resampling method together with a variety of performance metrics. Lastly, the blending ensemble models are presented. The final stage demonstrates the third study objective by experimenting upon the blending ensemble algorithm developed in this study in order to evaluate its efficacy. A series of experiments will be conducted to demonstrate the efficiency, and to appraise the performance of the blending ensemble model

against other machine learning models by applying a wide range of well-known measurable statistical metrics.

1.6 Structure of the Dissertation

The study is presented in five chapters, which are arranged in the following manner:

Chapter One: Introduction and Background

The first chapter introduces the background related to the study and highlights the issues and challenges encountered when dealing with fake news. This motivates for the necessity of developing an improved solution towards the detection of fake news. In addition, the research problem is clarified, which lead to the formulation of the research question, aim and objectives. The contribution of this study is highlighted in this chapter as well as the scope of the study.

Chapter Two: Literature Review

Chapter Two presents an extensive review of relevant publications that focus on the challenge of detecting fake news and the existing approaches to overcoming this challenge. This serves as a basis that underpins the study reported in this dissertation. This chapter provides a detailed account of topics associated directly with the study with an emphasis on evaluationg the strengths and shortcomings of the different methods and strategies for fake news detection as proposed in the literature.

Chapter Three: Research Methodology

The third chapter presents the major contribution of this study. It comprises the detailed step-by-step methodology carried out to accomplish the set research aim and objectives proposed in this study. Design research is introduced and applied to the study. The proposed framework is presented in detail together with the benchmark fake news datasets, performance metrics and resampling method. Finally, the blending ensembles are expounded upon.

Chapter Four: Presentation and Discussion of Results

Chapter Four provides a quantitative performance evaluation of the proposed blending ensemble algorithm results, analysis and comparison with other machine learning models. The experiments were conducted using two benchmark fake news datasets, and included GloVe and n-grams as features. Six well-known statistical metrics were applied during the evaluation of the models. Moreover, ROC Curves, P-R Curves and confusion matrices are also included in this chapter to visually enhance the interpretation and evaluation of the study results.

Chapter Five: Summary and Conclusions

Finally, Chapter Five presents an overview of the research conducted and concludes this dissertation with deliberations on possible research recommendations and extensions, and suggestions for future studies in the field.

1.7 Chapter Summary

This chapter has contextualised the study by positioning it within the landscape of social challenges created by fake news dissemination, thereby motivating for the necessity of developing an effective solution utilising an ensemble approach. The rationale and research objectives for the study were also elucidated. A comprehensive review of literature that serves as a framework for guiding the study follows in Chapter Two.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

The second chapter introduces a comprehensive review of related studies that have impacted upon the direction and focus of this study. The chapter comprises seven major sections. The first section provides a background to fake news. The next two sections will place emphasis on the classification and varieties of fake news. The fourth and fifth sections will highlight and review the strategies currently used to detect fake news. Finally, the detection algorithms and meta-analysis of fake news detection algorithms will be presented.

2.2 Background

In the past, information and news were primarily sourced from newspapers and television (Bondielli & Marcelloni, 2019). With the advent of the internet and social media however, 'news' has become much more easily accessible (Granik & Mesyura, 2017). Although the evolution of social networks is of great benefit to humanity, they can also be argued to have adversely impacted upon our existence as a global society. This evolution has brought about a harmful variant in news known as 'fake news' that enables misinformation to proliferate into all mediums (Collins *et al*., 2020). The "origins of fake news date back to before the printing press" (Burkhardt, 2017, p. 1) but this phenomenon has gained prominence in recent times (Klyuev, 2018).

From the uprisings, anti-government protests, and armed rebellions of the Arab Spring to the 2016 presidential election in the United States, 'fake news' has continually been in the midst of controversy the world over (Rampersad & Althiyabi, 2020). Wardle and Derakhshan (2017) contend that the ramifications of fabricated content and rumours in social media promote uncertainty and mistrust. Social media has become a central part of many people's lives, enabling users to share their thoughts and feelings, to access news, and to network with one another. Views, attitudes and sentiments are reflected effortlessly by user engagement and interaction. Therefore, information is effortlessly spread further via social networks. The convenience and minimal cost associated with using social networks draws together "collective intelligence", but concurrently results in a negative side effect in the form of the dissemination of misinformation by way of fake news (Lu & Li, 2020, p. 1).

Consequently, governments throughout the world are endeavouring to combat this problem (Roy *et al*., 2018). Some governments have become very perturbed about 'fake news' and this has resulted in the introduction of punitive legislation. For example, Germany has passed a law (the Netzwerkdurchsetzungsgesetz) with fines up to 50 million euros if online platforms do not eliminate unlawful content (Mayer, 2018). This encompasses both hate speech and 'fake news'.

The United Kingdom's Home Secretary indicated that law enforcement and government intelligence should be granted accessibilty to the encrypted message services such as Facebook Messenger (Sparrow, 2017). The motivation for this access was based on the necessity to fight the prolific diffusion of misinformation, which may cause harm to the public or impact on national security. This in turn could generate political polarisation and a greater intake of news that comes from biased media environment (Qiao *et al*., 2020).

People are seemingly oblivious to recognising deception, and many assume that the majority of information they encounter in any form is trustworthy and factual. Some have a tendency to be unconscionably sensitive towards knowledge that they do not comprehend (Pennycook *et al*., 2015). Furthermore, confirmation bias influences individuals to grasp only what they desire and what relates specifically to them (Rubin *et al*., 2016). Hence, the propagation and proliferation of false news is of dire concern since it has the capability of generating devastating repercussions.

It is argued that the pervasiveness of fake news in social media gives rise to a decline in the integrity and crediblity of the news, social aplomb, and political polarisation (Allcott & Gentzkow, 2017). Automatic fake news detection has now evolved into a critical challenge due to the escalation of diverse content (Shu *et al*., 2017; Figueira & Oliveira, 2017). Consequently, there is an urgent need to address the escalating problem in order to detect deceptive information as quickly and efficiently as possible.

2.3 Classification of Fake News

Fake news has existed since time immemorial but only started to disseminate broadly once the printing press was invented in 1439 (Biyani *et al*., 2016). Current research on fake news has cast the spotlight predominantly on classification of posts on social media and online news. Diverse

approaches have been suggested by many researchers with the objective of detecting deception. Nevertheless, there exists no common or universal definition for fake news (Khan *et al*., 2019).

Fake news can however be classified into a number of categories. Rubin *et al*. (2015) propose three distinct varieties of fake news representing erroneous or disingenuous reporting: (1) Large-scale Hoaxes; (2) Serious Fabrications; and (3) Humorous Fake Stories. Large-scale hoaxes comprise untruthful information masquerading as factual. Serious fabrications are news articles designed with malevolent intentions. Humorous fake stories are presented to entertain the reader, such as satirical articles disguised as news. Hence, fake news can be characterised by its form, intent and verification as being partially or totally untrue (Bondielli & Marcelloni, 2019). Regardless of its origin and/or intent, misinformation is very obviously the common thread that runs through all varieties of fake news as illustrated in the following table.

Table 2.1: Misinformation Classification Matrix (Thota *et al*., 2018).

	Satire or Parody	False Connection	Misleading Content	False Context	Imposter Content	Manipulated Content	Fabricated Content
Poor Journalism		✓	✓	✓			
To Parody	✓				✓		✓
To Provoke or to "punk"					✓	✓	✓
Passion				✓			
Partisanship			✓	✓			
Profit		✓			✓		✓
Political Influence			✓	✓		✓	✓
Propaganda			✓	✓	✓	✓	✓

Current studies habitually link fake news to “deceptive news, false news, satire news, disinformation, misinformation, cherry-picking, clickbait and rumour” (Zhou *et al*., 2019, p. 7). Specifically, “fake news is a news article that is intentionally and verifiably false” and may possibly deceive the reader (Allcott & Gentzkow, 2017, p. 4). This far more restrictive definition ensures that there is no ambiguity between fake news and any other associated concept.

2.4 Types of Fake News

There is little consensus when it comes to ascertaining the various types of false information. However, for the purposes of assessing content online, ten broad types of deceptive or false news have been identified and will be explained in the sections below (Webwise, 2021). For the purposes of developing fake news detection strategies, it is crucial to differentiate between the many types of bogus news formats that will be explored below in the following order: (1) Clickbait; (2) Propaganda; (3) Satire and Parody; (4) Hoaxes; (5) Sloppy Journalism; (6) Misleading Headings; (7) Biased/Slanted News; (8) Conspiracy Theories; (9) Rumour Mills; and (10) Hate News (Collins *et al*., 2020).

2.4.1 Clickbait

These are purposefully contrived stories intended to increase website visitors and thus gain revenue from advertisements for the websites. Clickbait exploits sensational headings to capture interest, and influence the clicks through to a publisher's website, usually at the cost of veracity or truthfulness (Campan *et al*., 2017). Biyani *et al*. (2016) mention eight varieties of clickbait namely (1) Exaggeration; (2) Inflammatory; (3) Teasing; (4) Bait-and-Switch; (5) Graphic; (6) Formatting; (7) Wrong; and (8) Ambiguous, all of which usually have misleading information (gossip) that is not connected to any headline. Nevertheless, clickbaits have unfortunately proved to be a thriving and profitable scheme. For example, the headline in Figure 2.1 below can be argued to be inflammatory, exaggerated, ambiguous and wrong.

Figure 2.1: CNN Headline (Cohen, 2014).

Since readers are fearful of this potential method of virus transmission, out of ignorance, fear or curiosity, they will most likely click on the link, which will take them to the publisher's website thereby generating revenue for that publisher. They may likely believe this fake news story and will spread this false news that may lead to panic, only to later discover what the WHO (World Health Organisation) actually states about the virus mutation, which is only speculation.

2.4.2 Propaganda

These are stories that are fashioned to consciously mislead the audience, encouraging a biased viewpoint or a particular political agenda. Tandoc *et al*. (2018, p. 11) state that these are "news stories which are created by a political entity to influence public perceptions". Propaganda is usually misleading or prejudiced in some manner as it does not express the complete truth due to of the necessity to promote a particular opinion, cause, person or product. Examples include the commercials and political signs that promote a candidate's view over another to gain more votes in the political race. The poster below further exemplifies the idea.

Figure 2.2: Propaganda Poster (YourDictionary, 2021).

2.4.3 Satire and Parody

These are executed through a fictitious story that overstates the facts from conventional media by employing comedy (Brummette *et al*., 2018). Numerous websites and social media accounts present bogus news as entertainment or parody. For example, The Daily Mash, Waterfod Whispers, The Onion, etc. Parody imitates a particular artist, writer or genre, using deliberate exaggeration to produce a comical effect. This is achieved by overstressing and replicating obvious features and is often confused with satire.

Although satire can be developed from parody there is however a difference between the two. Parody mimics the subject directly whilst satire pokes fun at a person/subject using indirect imitation. Furthermore, satire's objective is to correct the inadequacies in our society by employing criticism, as illustrated in the satirical cartoon.

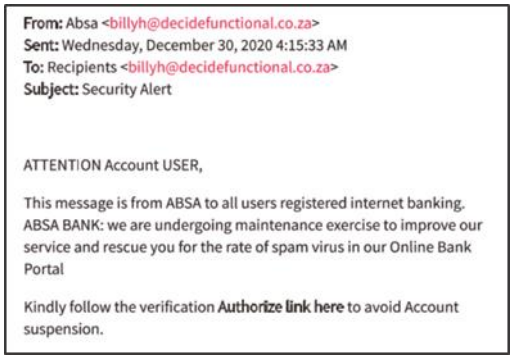
Figure 2.3: Satirical Cartoon (YourDictionary, 2021).

This graphic illustration depicts what we view on television but manipulates what we see in some manner to give the wrong message. In this cartoon, the man's poster states "Down with Police Brutality". However, it is showing a television news reporter with a pair of scissors and the word brutality lying on the ground so that the poster being filmed now states "Down with Police", which has a completely different meaning. Parody is used here to highlight the biased views of Fox News.

2.4.4 Hoaxes

Hoaxes are deliberately concocted stories created to mislead the public or audience. They are usually directed at public figures. Sometimes the mainstream media believe it to be factual and report it, which in turn often results in significant material damage to the victim (Rubin *et al*., 2015). The TextRank algorithm (Tammam *et al*., 2018) derived from the PageRank algorithm detects Indonesian hoax news with impressive results (Collins *et al*., 2020). Social media and emails are often used in the circulation of hoaxes. These may be used to trick the recipient into supplying personal login details to a phishing site. Once the victim falls for this scam the loss could be catastrophic, especially if the login credentials of a bank account have been solicited. The hoax message in Figure 2.4 below was sent to ABSA Bank clients as a security alert containing important "banking news" for this fraudulent purpose.

Figure 2.4: Hoax (ABSA, 2021).



From: Absa <billyh@decidefunctional.co.za>
Sent: Wednesday, December 30, 2020 4:15:33 AM
To: Recipients <billyh@decidefunctional.co.za>
Subject: Security Alert

ATTENTION Account USER,

This message is from ABSA to all users registered internet banking. ABSA BANK: we are undergoing maintenance exercise to improve our service and rescue you for the rate of spam virus in our Online Bank Portal

Kindly follow the verification **Authorize link here** to avoid Account suspension.

2.4.5 Sloppy Journalism

Digital media presents risks, opportunities and challenges to journalism. Although digitisation (internet) expedites the gathering of news and dissemination, it does not automatically improve the quality of journalism. Inadequate verification, plagiarism and unscrupulous journalistic practices have escalated at an alarming rate in many countries. Specifically, in developing nations such as South Africa, China, India, Brazil, Egypt, as well as in some developed countries like the UK, Finland and the US, there has been a decline in content originality and an escalation in "copy and paste" journalism. Furthermore, deadline pressures and the increased pace of journalism have also contributed to journalists becoming more susceptible to mistakes (Zannettou *et al*., 2019).

Occasionally, journalists or reporters may publish an article using unreliable material or without verifying the facts, which consequently leads to audiences being misimformed. For example, Urban Outfitters in the US issued an Election Day Guide, which incorrectly stated that voters needed a 'voter registration card'. However, this was not a requirement in any state for the US elections. More recently, Pretoria News in SA ran a story about a woman giving birth to decuplets on 7 June 2021. This record-breaking story spread the world over. However, about a week later it was confirmed as being "fake", and described as "a story of incompetence, negligence, abuse, and a cover-up" (Iol, 2021). Sloppy journalism and unreliable sources can adversely impact on public trust. For example, after the school shooting at Stoneman Douglas High in Florida, devious information went viral resulting in the public distrusting the news outlets (Bickham *et al*., 2018).

2.4.6 Misleading Headings

These stories are not entirely untruthful yet they can unfortunately be misrepresented by including sensational or misleading headlines. This type of news usually proliferates online since only small snippets and headlines are displayed on newsfeeds (Webwise, 2021). Misleading headlines used to capture public attention affect people's trust in the media since they can manufacture or exaggerate stories to distort the truth in the hopes of boosting the sales, subscriptions and traffic to websites. The truth is skewed by headlines that use quotes that are placed out of context and as such serve as an injustice to their readers. Consequently, it becomes challenging for readers to differentiate between news that is genuine versus opinionated or entertainment news. "The inability to decipher real news vs. fake news continues to decay trust in the media, public institutions and nonprofits" (Bickham *et al*., 2018, p. 10).

It is also challenging for reporters to compress a news story into a single heading using a very limited number of words. During this process, important information could be lost, which may result erroneously in the expression of something entirely different. A misleading caption could refer to an entire nation. For example, a headline carried by The Independent in May last year indicated that: "France sees 70 cases linked to schools days after reopening". During that period, schools were resuming in Europe and the UK, and anxiety increased as a result of this headline, which may in turn have lead to an escalation in Covid-19 cases (Full Fact, 2021).

2.4.7 Biased/Slanted News

A great deal of people are attracted to stories or news that resonate with their individual biases or personal beliefs. Bogus news can unfortunately feed on those biases. Newsfeeds are likely to present articles and news depending on our search history (Webwise, 2021). These are stories that are particularly biased or one-sided. From a political standpoint, they are branded as hyper partisan news (Potthast *et al*., 2017) and are extremely biased towards a party/person/event/situation.

Biased news has become quite problematic for some communities. Some Muslim, Hispanic and African American minorities believe that the media disregards their plights and frequently represent them in an unfavourable light. Biased or slanted reports have depicted Muslims as extremists, African Americans as criminals and Mexicans as job stealers. A while back, the

Washington Post published a Twitter story about the Russians who had hacked into an electrical power grid in Vermont. The state's governor named Vladimir Putin a "thug," and Senator Patrick Leahy said, "That is a direct threat to Vermont, and we do not take it lightly." However, two days later, The Washington Post retracted the story but sadly the libel damage had already been done. These prejudicial portrayals are often the cause of mistrust and conflict in the public media. To conquer public trust and support, it is imperative that slanted or biased news is highlighted and its message underscored (Bickham *et al*., 2018).

2.4.8 Conspiracy Theory

Sources that promote strange or eccentric conspiracy theories attempt to clarify an event or situation by calling on a popular conspiracy with no evidence of it being true. By in large, these stories involve illegal acts carried out by powerful individuals or governments. Unsourced information is presented as fact (Zannettou *et al*., 2019). Recent examples include the so-called hypothesis that COVID-19 was produced by a laboratory in Wuhan, and that 5G networks are triggering or intensifying COVID-19 symptoms (van der Linden *et al*., 2020).

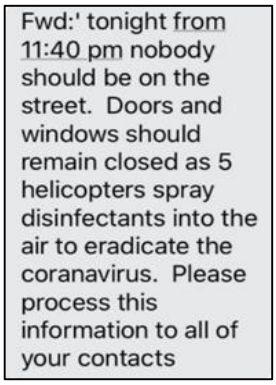
More importantly, this kind of misinformation has had other serious societal repercussions. Recent reports indicate that COVID-19 misinformation is linked to mass poisonings and mob attacks. People have also burnt at least fifty cell phone masts in reaction to the 5G conspiracy (BBC, 2020).

2.4.9 Rumour Mills

These are sources that deal in gossip, rumours, unverified claims and innuendo. The Oxford English Dictionary defines rumour as "a currently circulating story or report of uncertain or doubtful truth". Hence, the unverified content could prove to be either correct, or partially or completely untrue, or it could continue to be unresolved. The growing use of and dependence on social media networks for news and correspondence and the absence of monitoring often promotes the surfacing and growth of rumours that are, "unverified at the time of posting" (Zubiaga *et al*., 2018, p. 1). These authors introduce two categories of rumours: "long-standing" and "newly emerging" rumours. The former is in circulation for a prolonged time, whilst the latter usually sprout with breaking news, where stories may be disjointed and often unconfirmed at the

beginning. The rumour displayed in Figure 2.5 below was posted in March 2020 on Facebook and WhatsApp, and escalated to viral status in no time.

Figure 2.5: Facebook Rumour (BBC, 2020).



The individual behind the post failed to clarify how the aerial sprayings would eradicate the coronavirus indoors, where it is likely to spread the most. Nevertheless, the post stated to stay indoors with windows and doors closed. This message was shared in the UK and spread to Pakistan, Spain, Colombia, the Netherlands and beyond, even though the time and number of helicopters remained the same. The result: A nonsensical message was taken seriously by many.

2.4.10 Hate News

These are stories that actively promote homophobia, misogyny, racism and all other forms of discrimination. Hate and bigotry are sadly nothing new however social media and websites have intensified their force since the web is now the primary source of news aquisition. Furthermore, news discussions now involve more social interactions with users posting commentaries or discussing news on various online platforms such as WhatsApp (Zannettou *et al*., 2020).

Anyone with access to a cell phone and data can now share and debate news articles with relative ease from the comfort and safety of their own home, which tends to yield inflammatory and unfounded statements that can promote discrimination (Zannettou *et al*., 2017). So although the capability to post remarks online empowers users to deliberate freely over news, discussions can also turn toxic, triggering hate speech and racist remarks (Harlow, 2015). On January 6, 2021, a mob of rioters attempted an insurgency on the US Capitol. Social media played a major role in

attracting the participants to Washington, DC to actively participate in the action. These included members belonging to white supremacist hate groups who came forward challenging Joe Biden's victory in the 2021 US presidential elections. The events that unfolded demonstrate how online hate can ignite unnecessary savagery and violence (Science News for Students, 2021).

Recent studies show that polarised communities on the Web regularly organise and coordinate campaigns, and members are ordered to "attack" certain targets with hate speech (Hine *et al*., 2017; Flores-Saviaga *et al*., 2018). "Raids" can sometimes be aimed at news originating from websites that promote agendas that are in conflict with these members. Notwithstanding the problem of news comments containing hate speech, moderation of commentaries continues to be a prevailing problem (Pavlopoulos *et al*., 2017). For this reason, hate speech on social media platforms has been studied by a several researchers (Davidson *et al*., 2017; ElSherief *et al*., 2018).

2.5 Detection of Fake News

Fact-checking involves assessing the truthfulness or veracity of a claim. The task is usually carried out manually by journalists who verify the claims of public figures (Vlachos & Riedel, 2014). In addition, ordinary citizens are also expected to carry out fact-checking on the voluminous content (statements) that they consume. The consequence of social media platforms, misleading of the public by fake news, concealment of the truth and the repercussions thereof have necessitated the battle against fake news (Ünal & Çiçeklioğlu, 2019). To this end, fact-checking (verification) platforms have come to the fore to verify information that has been dispersed via traditional media as well as online and particularly on social media platforms. These fact-checking platforms can be viewed as service providers that validate the claims of public statements by investigating both the primary and secondary sources (Brandtzaeg & Følstad, 2017). Fact-checking systems can therefore be quite useful to society by aiding in stemming the flow of fake news and are in fact considered "as a potential tool to combat fake news" (Chung & Kim, 2021, p. 3).

2.5.1 Manual Fact-checking

This process is conducted by humans who are either ordinary citizens or experts. It is a time-consuming and cumbersome task. There is often a significant length of time that passes from the moment a claim (statement) is made to the publication of the fact-check (Hassan *et al*., 2015). The

slow progress of the checking process life cycle is thus hindered by the wearisome nature of the task. A lot of time is consumed in identifying claims for checking. Journalists need to sacrifice hours in examining transcripts of interviews, debates and speeches to isolate claims for further research (Popat *et al*., 2017). Furthermore, advanced research skills are essential in fact-checking.

2.5.2 Expert- based Fact-checking

This method relies on professionals in the field of fact-checking, also known as fact-checkers, to validate the content of a specific news item via manual research. During the process, specific components are assigned certainties when the text's accuracy is compared to others that have already been checked (Vlachos & Riedel, 2014). This is normally carried out by a small number of trustworthy fact-checkers, and it is crucial that the documents and data are thoroughly scrutinised (Ahmed *et al*., 2019). This approach is accurate however the main challenge is the high cost incurred by hiring professionals. Furthermore, the system is vulnerable to overburdening as the volume of the content for authentication escalates (Ciampaglia *et al*., 2015).

To assist in this regard, there exist fact-checking organisations and projects, which play an important role. FactCheck.org, based at the University of Pennsylvania, "aims to reduce the level of deception and confusion in US politics" (FactCheck, 2021). PolitiFact is an organisation owned by the non-profit Poynter Institute for Media Studies. It's primary focus is political fact-checking and is operated by journalists. David Mikkelson founded the Snopes.com website that began as a project in 1994 in researching "urban legends", and which has now become the largest and oldest fact-checking website on the net to date (LibGuides, 2021).

2.5.3 Crowd-sourced Fact-checking

This is an alternative variation of fact-checking that involves a large group of ordinary individuals who perform the function of checkers. Crowdsourcing allows the group (crowd) of individuals to reach a collaborative judgement by investigating the authenticity of news (Pennycook & Rand, 2019). Therefore, the accuracy of the news is entirely established by the knowledge present in the crowd (Hassan *et al*., 2017; Ahmed *et al*., 2019).

This type of fact verification is not simple to perform, and the outcome is expected to be less accurate and less trustworthy due to the individual prejudices of the fact-checkers. On the plus side, it is unlikely for this system to become bogged down when a surge in content has to be verified. It is essential to filter out untrustworthy users and sort out any contrasting results. These issues become increasingly pertinent as the cluster of fact-checkers begins to grow. Nonetheless, individuals involved in conducting fact-checking through crowdsourcing are capable of delivering an abundance of feedback such as their sentiments, opinions and attitudes (Hassan *et al*., 2019).

2.5.4 Automatic Fact checking

The aim here is to provide an automatic fact-checking method to users that is capable of identifying whether a particular news item is false or true (Ahmed *et al*., 2019). A major difficulty with human fact-checking arises when the system is overwhelmed with increasing volumes of current news to be verified, which is the norm on social media (Rosenkilde, 2017). Therefore, methods involving automated procedures have been designed to avoid this eventuality. These approaches generally rely on Natural Language Processing, Information Retrieval Techniques or Network/Graph Theory (Zhou *et al.*, 2019).

Automatic fact-checking techniques normally involve two steps; fact-extraction and fact-checking. Fact-extraction grasps knowledge, or "raw facts", from the web that appear characteristically conflicting, unnecessary, obsolete, incomplete or inaccurate (Nickel *et al*., 2015). Knowledge processing tasks are used to clean up and refine the knowledge in order to create a knowledge graph or knowledge base. Fact-checking is then applied to assess the news content in order to establish authenticity. This process is performed by matching or linking the news that requires checking to those evidences present in the knowledge base(s).

2.6 Detection Strategies for Fake News

The detection of fake news is a complicated undertaking, largely due to its inherent attributes and nature (Ahmed *et al*., 2017). Detection strategies are therefore used to take advantage of several news-related types such as publisher, headline, body, as well as social-related types such as spreaders, propagation paths and feedback (Zhou *et al*., 2019).

The methods used in the detection of fake news are usually divided into content and context (social) based methods (Shu *et al*., 2017). The distinguishing feature between these two approaches is whether or not they depend on information related to the social context. This refers to information about the propagation of news on social media and includes the ancillary information associated with subscribers of social media and their networks (connections) (Zhou *et al*., 2020).

Ruchansky *et al*. (2017) and Della Vedova *et al*. (2018) have presented solutions to assess social context information. The detection of fake news improves with accessibility of additional social context information. As such, it becomes very challenging to detect fake news that has just been published and not yet propagated. The news content itself also has a significant role to play during the fake news detection process. Current content-based detection methods utilise text information or a combination of the two data types (Jin *et al*., 2016; Wang *et al*., 2018; Yang *et al*., 2018,).

Content-based methods focus on the news' content itself, such as the title, body text and social signals and features, for example, users' interactions related to news posted on social media in the form of "liking" an item on Facebook or "retweeting" an item using Twitter, etc. An analogous categorisation of methods has previously been suggested by Conroy *et al*. (2015), who furthermore indicated that machine learning should be integrated with these approaches.

Content-based methods have been employed in traditional news media, and generally in instances where social data is absent. These methods have historically been applied in detecting junk email but more recently have been employed in detecting fake news. Reidel *et al*. (2017) demonstrated by utilising a fairly straightforward approach including term frequency combined with inverse document frequency that yielded an 88.5% accuracy. Ahmed *et al*. (2017) exploited a similar approach on six distinct classifiers obtaining a 92% accuracy on a two thousand news dataset.

In summary, the task of identifying false information can be tackled manually and by using automated methods to verify the authenticity of information (Jin *et al*., 2016). Manual efforts rely on crowd knowledge or experts. Users of social media can be engaged by requests to flag possible fake content, which can then be investigated further. Naturally, manual detection is time

consuming and impractical, attributable to the substantial amount of content spawned on social media. Automated methods are thus deemed more appropriate for efficient detection of fake news.

2.6.1 Knowledge-based

Recent studies support the amalgamation of knowledge engineering and machine learning in detecting fake news (Hinkelmann *et al*., 2019). However, the rapid pace of the dissemination of fake news on social media poses a demanding challenge. Platforms like Twitter enable little pieces of fabricated information to spread speedily (de Beer & Matthee, 2020). Knowledge-based approaches strive to use external sources to verify the veracity of the news before the spread escalates.

Fact-checking is applied during knowledge-based fake news detection. As previously mentioned, fact-checking has its roots in journalism and assesses the authenticity of news by comparing its content (statements and claims) with the acknowledged facts. The two categories applied are manual (traditional) and automated checking of facts (Zhou & Zafarani, 2020).

2.6.2 Style-based

Style-based and knowledge-based detection are similar in nature to a degree, since both approaches focus on the content analysis of news. Knowledge-based techniques primarily assess the truthfulness of news, whilst style-based evaluate the purpose of the news. The question posed is whether there is intention to deceive the public? The supposition premised on style-based approaches is that entities that are potentially harmful generally use a “special” style or technique in writing fake news to persuade others to read and influence them to believe (Zhou *et al*. 2019).

Both knowledge-based and style-based fake news detection methods rely on how efficiently the news content style is captured and represented, and the classifier’s performance on different representations of the news content. Usually a feature vector is used to represent the content style in a machine learning context to establish if the news content is untrustworthy.

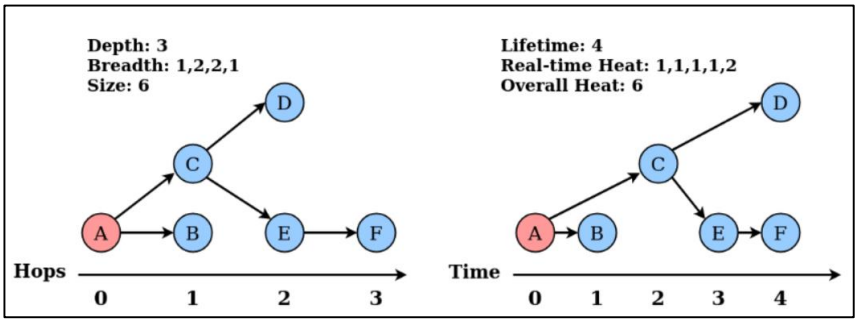
2.6.3 Propagation-based

Propagation-based detection determines the trustworthiness of the sources of news at different stages, from creation to publishing and the circulation by social media, by analysing the spread of fake news (Qiao *et al*., 2020). These include network-based and cascade-based methods which are predominately used to detect fake news.

2.6.4 Cascade-based

A cascade for fake news is represented using a tree-like structure that depicts how the fake news has been propagated by users on social media. The node at the root indicates the user who created the fake content. The remainder of the nodes denote those users who thereafter propagated the news by reposting or forwarding. The cascade is expressed either by the steps travelled by fake news, referred to as "Hops-based Fake News Cascade", or by the number of repostings, referred to as "Time-based Fake News Cascade" (de Oliveira *et al*., 2021). The Hops Cascade is frequently denoted using a standard tree including parameters such as size, breadth and depth. Depth depicts the greatest number of hops taken whilst breadth numerically represents all the users who received the post. The predominately size equals the total user number of the cascade. Similarly, a time-based cascade can be used but with parameters termed lifetime and "heat" which is seperated into two categories of heat. Lifetime is the lengthiest interval for the propagation, real-time heat refers to the number of users who have forwarded and reposted the content at time (t), and overall heat indicates the all-inclusive user number who have reposted or forwarded the fake content. Figure 2.6 below illustrates the relationship between the two news cascades.

Figure 2.6: News Cascades (de Oliveira *et al*., 2021, p. 20).



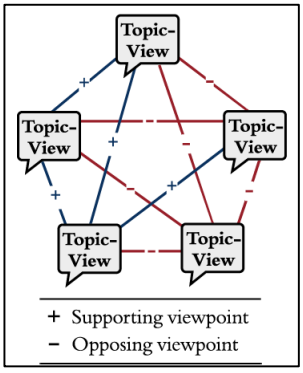
2.6.5 Network-based

With this method, flexible networks are constructed to indirectly capture how fake news is propagated. The networks used are either hierarchical, homogeneous or heterogeneous as will be elaborated upon in the sub-sections below.

2.6.5.1 Homogeneous network

These networks contain one type of node and one type of edge as illustrated in Figure 2.7 below. In the figure, a stance network is shown where the nodes denote the news-related post of the user and the edges denote the negative (opposing viewpoint) or positive relation (supporting viewpoint) between posts. The truthfulness of news posts can be evaluated by using such a network.

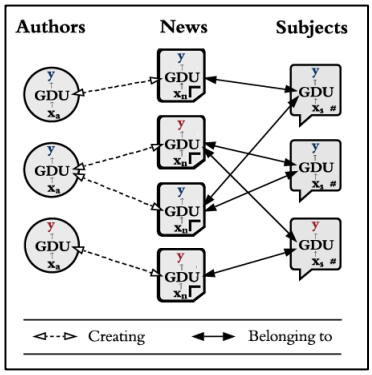
Figure 2.7: Homogeneous Network (Zhou *et al.*, 2019, p. 23).



2.6.5.2 Heterogeneous Networks

Networks of this nature comprise multiple types of edges and nodes. It usually takes the form of a framework consisting of three components, namely: (1) entity embedding and representation; (2) modelling of the association; and (3) semi-supervised learning. An example, shown in Figure 2.8 below, illustrates the relationship network between authors, news, and subjects. The authors are the creators of the news and subjects are the key focus in the news.

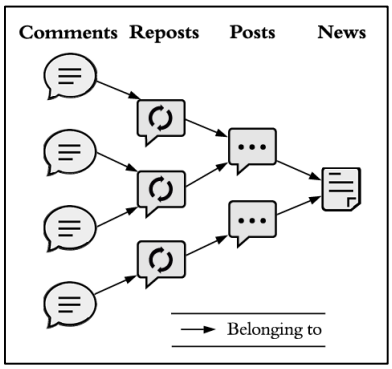
Figure 2.8: Heterogeneous Network (Zhou *et al*. 2019, p. 23).



2.6.5.3 Hierarchical Networks

This network comprises edges and nodes from a variety of types that form a hierarchical relationship (set/subset). News authentication can be converted to a problem requiring graph optimization using a hierarchical network. An illustration is the news, tweet, retweet, reply network shown below in Figure 2.9.

Figure 2.9: Hierarchical Network (Zhou *et al.* 2019, p. 23).



2.6.6 Source-based

Bogus news is likely best identified by evaluating the 'credibility' of its source. Credibility is frequently defined in terms of quality and believability (Viviani & Pasi, 2017), "offering reasonable grounds for being believed" (Zhou *et al.* 2019, p. 26). This approach examines fake news based on information related to the news and its social context. For example, an article originating from an unreliable website, then distributed by an untrustworthy user is possibly fake when compared to a report sent out by a credible or respected user. Identification of the news

content's source is thus key to the success of this approach, however the challenge is that originators of fake news are becoming increasingly adept at covering their tracks online. Four interlinked approaches to source-based detection are outlined below.

2.6.6.1 News Headline Credibility Assessment

This method involves detecting clickbaits, which are headings used to entice users to click on a particular web page link. Both linguistic and non-linguistic features have been used to detect clickbait. Linguistic features such as term frequencies, forward references and readability, and non-linguistic features such as headline stance, webpage links, and user interests have all been used to identify clickbait within a "supervised" learning context (Biyani *et al*., 2016).

2.6.6.2 News Source Credibility Assessment

This approach examines the political bias, credibility and calibre of the source website with the intention of determing the reliability and quality of the news content. As mentioned however, determining the correct source is often difficult.

2.6.6.3 News comments credibility assessment

The trustworthiness of news can also be assessed by examining the associated user comments in terms of credibility as they may contain valuable information on opinions and stances. Content, behaviour and graph(network)-based models can also be used to assess credibility of comments.

2.6.6.4 News Spreader Credibility Assessment

News content credibility can be assessed by investigating the users that propagate the content and evaluating their trustworthiness. Users are an essential component in disseminating misleading information. Fake news is effortlessly spread through sharing, liking, forwarding and reviewing. Users are categorised as either malicious (low reliability) or normal (higher reliability) (Zhou *et al*., 2019). Users that are malicious intentionally circulate false news in the quest for financial gains, popularity or power. Malicious users include automated bots, trolls and cyborgs.

2.6.7 Language-based

The language-based detection strategy draws on conventional linguistics, style-associated readability features, or word embedding, to differentiate real from fake news (Qiaoet *et al*., 2020). This approach applies linguistics either through software or humans to detect fake news (de Beer & Matthee, 2020). The spreaders of misinformation possess power over their content primarily due to their anonymity, but their identities can be revealed by the style present in their language use (Yang *et al*., 2018). Attention is therefore focused on syntax and grammar. Hence, every word, the characters in the word, their structure and how all comes together to form a paragraph are analysed for patterns (Burkhardt, 2017). The main methods applied in the language approach are Bag-of-Words, deep syntax analysis, and semantic analysis (de Beer & Matthee, 2020).

In Bag-of-Words (BoW) the individual word frequencies are examined to uncover indications of misinformation. All words are regarded as independent entities and are assigned equal importance (Burkhardt, 2017). This eventually aids in identifying word usage patterns that assist in detecting deceptive information. This model however overlooks the context once text is transformed to a numerical representation, and the location of words is often not considered (Potthast *et al*., 2017).

Deep Syntax Probability Context Free Grammars are used to implement the deep syntax method. This is accomplished by utilising parse trees that enable the analysis of Context Free Grammars. Probability Context Free Grammars can be considered as an expansion of Context Free Grammars. Rules are derived from sentences which are then applied to analyse the different syntax structures. By comparing the syntax to patterns or known structures can finally lead to differentiating between real news and fake news (Burkhardt, 2017).

Semantic analysis veracity can be resolved by associating personal experiences, such as a hotel review, to a profile of the topic obtained from related articles. A truthful writer is very likely to express comparable comments on a subject to other honest writers. The approach uses different compatibility scores (de Beer & Matthee, 2020).

In general, fake news detection methods based on language are more beneficial when compared to propagation and knowledge-based approaches, since they are scalable and allow for close real-time feedback (Qiao *et al*., 2020).

2.6.8 Topic-agnostic Approach

Topic-agnostic features are taken into consideration during the detection process rather than the news article content. A few examples of these features are pages that have a substantial number of advertisements, existence of an author name, and longer eye-catching headlines (Horne & Adali, 2017). This approach employs both web mark-up capabilities and linguistic features in order to determine fake news (Castelo *et al*., 2019).

2.7 Detection Algorithms

2.7.1 Classical Machine Learning Approaches

News content is expressed through the use of hand-picked feature sets that fall within a traditional/classical machine learning framework. These sets can include both latent and non-latent features and are derived from the textual content on various linguistic levels (lexicon, syntax, semantics, and discourse) or images (Pérez-Rosas *et al*., 2017).

Machine learning models may be unsupervised, semi-supervised, or supervised, and can effectively detect news using these representations. Supervised classifiers are commonly used in style-based detection of fake news. For example, these approaches have relied on Random Forests (RFs) (Zhou *et al*., 2020), and Support Vector Machines (SVMs) (Pérez-Rosas *et al*., 2017) and will be explored below.

Zhou and Zafarani (2020) revealed the following when forecasting fake news in a conventional machine learning framework: (1) latent features are frequently outperformed by non-latent features; (2) a combination of features from different levels can outclass choosing just a single-level feature; and (3) rewrite rules and lexicon frequencies that denote fake news content style deliver superior performance when compared to other feature groups, despite their relatively lengthy computation time. It must, however, be reinforced that classifiers execute their best results in settings originally intended for them (Fernández-Delgado *et al*., 2014).

Since initial attempts targeting deception detection yielded encouraging outcomes, machine learning approaches have been applied in numerous research studies that assess the problem of fake information detection. A supervised machine learning strategy was implemented in the majority of studies by using Support Vector Machines (Bondielli & Marcelloni 2019).

SVMs are discriminating classifiers that are defined by a dividing hyperplane. They have outclassed several supervised machine learning approaches in detecting deception. Content-based features, such as visual and linguistic features, have been taken advantage of with worthy effect in many SVM approaches to deception and fake news detection (Pérez-Rosas & Mihalcea, 2015; Rubin *et al*., 2016). In particular, Rubin *et al*. (2016) trained a SVM to detect satirical fake news by applying features based on content and attained 0.87 as an f1 score. SVM-based approaches have also adopted a combination of context and content features for the training of classifiers, which are then utilised to tag an event reported on numerous online posts as being a doubtful or otherwise (Qin *et al*., 2016). Furthermore, SVM approaches have also provided a group of features directed at detecting anything novel in tweets, and attained a 0.75 accuracy score.

An interesting SVM strategy has been executed by Wu *et al*. (2015). They proposed a SVM strategy based on a graph kernel for rumour detection. Content features and propagation structures were utilised in recording 0.91 for accuracy.

Horne and Adali (2017) furnished a set of features to distinguish real news from fake news and satire. The outcome of the experiment suggests that titles/headings are important for differentiating fake and real news, and that satire is comparable to fake news.

Chakrabortyet *et al*. (2016) further exploited SVMs for clickbait detection. The authors worked with a feature set derived from the content and obtained a 0.93 f1 score.

Another well studied and extensively used algorithm is the decision tree (Breiman *et al*., 2017), which implements a recursive split on the values of features with the objective of determining the class. Decision trees can be derived from data by algorithms , for example J48 (C4.5). They have depicted competitive performance in comparison to other machine learning algorithms despite

their simplicity. The efficacy of the J48 decision tree has been demonstrated with various algorithms (Zhao *et al*., 2015; Giasemidis *et al*., 2016; Aker *et al*., 2017).

A blend of context- and content-based features have been used to assess the credibility associated with tweets. Bodnar *et al*. (2014) suggest a sequence of metrics for representing the user's trust in order to appraise the honesty of social networks by means of decision trees. The authors achieved an accuracy score of 0.75. Zhao *et al*. (2015) used curious tweets to determine potential rumour clusters. The associated features describing each cluster were fed into a J48 decision tree producing a 0.52 accuracy score. Giasemidis *et al*. (2016) executed decision trees during reliability classification involving rumours. This approach yielded a 0.96 accuracy on a dataset consisting of a relatively small feature set.

Random Forests (RFs) have been successfully implemented in several works. Numerous decision trees are ensembled to form a random forest. The predictions made originating from the individual trees indicate the eventual result. Aker *et al*. (2017), Briscoe *et al*. (2014), and Giasemidis *et al*. (2016), have demonstrated that RFs are strong performers in relation to other machine learning algorithms involving tasks in fake news. Kwon *et al*. (2013) adopted a RF for classifying rumours in a tweet graph by using a set of structural, linguistic and temporal features, attaining 0.90 accuracy. Aker *et al*. (2017) and Zeng *et al*. (2016) exploited RF in the detection of stance. Zeng *et al*. (2016) applied a content-based approach with regards to tweets resulting in an average accuracy ranging from 0.83 to 0.88. Aker *et al*. (2017) proposed using features sourced from prior works (Hamidian & Diab, 2019; Zubiaga *et al*., 2016) together with a set of features based on stance. The authors recorded an accuracy of 0.79 using RumorEval (Derczynski *et al*., 2017). Credibility evaluation has also taken advantage of RFs with positive effect (Ito *et al*., 2015).

Learning algorithms established on logistic regression have also been used extensively in research related to rumour stance classification. Similar to RFs and decision trees, comparative studies involving diverse approaches to rumours and fake news have recorded favourable results in terms of logistic regression (Aker *et al*., 2017; Ballarin *et al*., 2017; Enayet & El-Beltagy, 2017; Giasemidis *et al*., 2016).

Ferreira and Vlachos (2016) made use of logistic regression to classify news articles on the stance between headlines and claims with 0.73 accuracy. Chua and Banerjee (2016) studied linguistic predictors associated with rumour veracity by employing logistic regression to identify predictors of importance with regard to rumour verification. Hardalov *et al*. (2016) deployed logistic regression in evaluating the trustworthiness of news from Bulgaria, attaining a 0.75 accuracy score.

2.7.1.1 Hybrid Approaches

As earlier explained, human beings are not very proficient at recognising fake news. According to Okoro *et al*. (2018), humans have a 54% success rate in identifying fake news, and this plummets to just 4% if they hazard a guess. By using a hybrid model on social media composed of machine learning as well as human learning the success rate for fake news detection was improved. The effectiveness of the model was accelerated by using a combination of news from social media, a network approach and machine learning. The model identified the probability of the news that might be false.

The CSI model (Ruchansky *et al*., 2017) is a hybrid model that **c**aptures, **s**cores and **i**ntegrates. A Recurrent Neutral Network (RNN) extracts article representations, then a score and representation vector is created, and finally the outputs from capture and score are integrated forming a vector that is used during classification.

2.7.1.2 Ensemble Approaches

With respect to machine learning and statistics, ensemble methods apply several learning algorithms in pursuit of achieving superior predictive performance that could be sourced from any of the individual component learning algorithms (Granik & Mesyura, 2017). In contrast to statistical ensembles, which are generally infinite, ensembles in machine learning comprise of a finite set of models.

The evaluation of an ensemble's prediction typically involves additional computation over and above that which is required for an individual model, hence ensembles can be viewed as a means of making up for learning algorithms that are weak by executing a considerable amount of additional computations. For example, decision tree algorithms are frequently included in

ensembles (eg. RFs) because of their speediness. Slower algorithms however can also benefit from ensemble learning techniques (Kowsari *et al*., 2018). Ensemble methods have been employed in unsupervised learning as well, for example in anomaly detection (Zhang *et al*., 2019) and in consensus clustering (Alguliyev *et al*., 2020).

Ensembles have a tendency to generate improved performance when there is substantial variety amongst the models. Therefore, various ensemble approaches aim to support diversity among the base models that are eventually combined. Additional random algorithms, such as random decision trees, may be included in order to yield a more robust ensemble than for example entropy-reducing decision trees. Utilising a combined assortment of strong learning algorithms has proven to be more effective than employing individual techniques that endeavour to simplify the models so as to promote diversity (Kowsari *et al*., 2018).

Agarwal and Dixit (2020) applied ensemble learning by combining numerous dissimilar models and utilising the base classifiers in order to increase approximation, prediction and classification of the model. The ensemble was built using a combination of Naive Bayes, k-nearest neighbours (KNN), SVM and deep neural networks models. The resultant classifier had the least error and possessed better predictive power than all the other models in the mix. This approach is advantageous since it can diminish a threat of those particularly underperforming classifiers by including numerous models and then using the average to obtain the final model (Ghosal *et al*., 2018).

Ensemble machine learning uses a group of classifiers where individual decisions are often aggregated using weighted voting to achieve improved predictions, and decrease bias and variance. Mohale and Leung (2019) presented an ensemble learning model to ascertain the truth probability of statements used in social networks by taking into consideration the related metadata. The system used five different classifiers to improve the identification of phony news, yielding average accuracies of 80%. Reddy *et al*. (2020) evaluated text mining methods to detect fake news in a hybrid combining stylometric features and vector representations using ensemble methods such as boosting, bagging and voting, attaining accuracies of up to 95.49%. Mahabub (2020) presented eleven familiar machine-learning algorithms to handle the classification of news. The top three

machine learning algorithms were selected to create an Ensemble Voting Classifier, which attained 94.5% accuracy. Ahmad *et al*. (2020) explored distinct textual properties that were used to differentiate real from fake content. By using those properties, different combinations of machine learning algorithms were ensembled and evaluated on four datasets. Experimental results confirmed the superior ensemble learner performance over individual learners. Al_Ash *et al*. (2020) employed several classifiers in majority voting ensemble. The ensemble learning approach helped to avoid overfitting due to high dimensional and small dataset size.

Traditional machine learning models have been both successfully and widely applied (Crawford *et al*., 2015). The typical machine learning approach uses a single learning model in problem solving. However, contemporary studies are validating the effectiveness of "ensemble learning". The capacity of weak classifiers may be enhanced by merging the outcomes of several classifiers resulting in a more "robust" classification (Gutierrez-Espinoza *et al*., 2020). Through voting (or averaging) of the outputs created by a pool of classifiers, particularly smaller datasets, ensemble techniques deliver improved predictions and simultaneously evade overfitting. Furthermore, by involving multiple models, the search space is increased and the probability of obtaining better outputs becomes greater (Sagi & Rokach, 2018). Ensemble learning may thereby help to overcome machine learning challenges and constraints that include class imbalance, dimensionality and concept drift, among others.

The main tactics for building an ensemble learner include stacking, bagging, and boosting. Boosting generally makes use of base models that are homogeneous and are trained in sequence; bagging is comparable to boosting but the training of the base models is done parallelly; and stacking employs mostly heterogeneous models as the base with training taking place in parallel. A meta-model is used for combining the base models (Gutierrez-Espinoza *et al*., 2020). Boosting and bagging are the most popular methods for ensemble learning. Both approaches introduce diversity through modifying the training set, in a manner such that the learning algorithm is carried out numerous times on different training sets. The key difference, however, is that bagging performs random sampling of the data using replacement, whereas boosting does the same but on weighted data. These weights are updated iteratively in an attempt to provide more importance to

previously misclassified samples. RF is a popular ensemble and a special form of bagging that is combined with tree models (Bolón-Canedo & Alonso-Betanzos, 2019).

Stacked generalisation (stacking) includes the training of a learning algorithm to amalgamate the predictions of a number of base learning algorithms. The available data is used in the training of all the algorithms. The final prediction is revealed by a "combiner algorithm" (Kowsari *et al*., 2018), that takes into acoount all the predictions by the base algorithms.

Blending is another tactic used for building an ensemble learning approach and is most comparable to stacking, and can in fact be classified as a variation of stacking. The variance is that predictions based on a holdout validation dataset are used to fit the meta-model, instead of predictions made out-of-fold. The blending model ultimately conflates the predictions of a number of base models in determining the final prediction. Studies involving the application of blending models to fake news detection are very difficult to come by and as such have been assessed by the researcher as virtually non-existent, thereby representing a gap in the field of fake news detection strategies.

2.7.2 Deep Learning Approaches

This particular area of machine learning approaches has been explored extensively and applied in a broad range of research applications. Deep learning classifiers have enjoyed an unparalleled rise to fame in modern times, attributable to very encouraging results in several fields of research, for example text mining and the processing of natural language (Bondielli & Marcelloni 2019). Deep learning frameworks have the ability to learn the hidden representations of simpler inputs including variations in both content and context. News content, comprising of images and text, is often embedded at a tensor or pixel matrix for images, and at a word-level for text. The embedding is subsequently processed by neutral networks that have been well trained such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) (Huang et al., 2017). The news articles are eventually classified as fake or true by feeding the concatenated features into the classifier. This can be improved by enhancing feature representativeness. The Event Adversarial Neural Network applied by Wang *et al*. (2018) enhances the representativeness of features through the extraction of features that are unvaried.

The focus of deep learning is on modelling the network rather than modelling the appropriate input features, thus enabling the task to be resolved more effectively (Ma *et al*., 2016). Conventional machine learning uses features that are handcrafted as representations. The task of extracting features can be laborious though and could produce a subjective set of features. Therefore, the recognition of pertinent features for investigation could lead to a larger challenge, which is crucial for tasks related to fake news (Bondielli & Marcelloni, 2019).

Neural networks such as CNN and RNN are extensively used for this purpose. The nodes in RNNs are connected in sequence to each other forming a directed graph, which enables RNNs to be very effective especially in the modelling of data of a sequential nature, for example language, and also in obtaining pertinent features from distinct sources (Ruchansky *et al*., 2017).

In terms of deception and rumours, one of the first implementations of RNNs applied during detection is presented in Ma *et al*. (2016). RNNs with different architectures are proposed in Cho *et al*. (2014). The Gated Recurrent Unit achieved the best results of 0.88 and 0.91 accuracy on the two datasets considered. Furthermore, traditional machine learning algorithms were comprehensively beaten by deep learning approaches.

CNNs comprise different layers, namely, input, output, and a number of hidden layers. The data is transformed by means of pooling and convolution processes. CNNs are regarded as "state-of-the-art" for numerous computer vision applications and have been extensively applied in the processing and recognition of images (LeCunet *et al*., 2010). They are also gaining traction in natural language processing too (Jacovi et al., 2018). Recent works on rumours and fake news implement the CNN architecture. Chen *et al*. (2017) applied a CNN together with word embeddings for veracity and classification of stance involving tweets. The authors achieved accuracy scores of 0.53 and 0.70 respectively. Yu *et al*. (2017) exploited CNNs and paragraph embeddings in learning how posts are represented related to a particular event and then fed them into a CNN model. The strategy yielded an accuracy score of 0.93. Volkova *et al*. (2017) assessed both a CNN and an RNN approach in the identification of trusted and suspicious (e.g. hoaxes, propaganda) posts associated with news. The architecture implemented sequences of words

together with network and linguistic cues related to deception. The average precision score of almost 1.00 was reported on both approaches.

Ruchansky *et al*. (2017) employed Long Short Term Memory (LSTM) in a study to combat fake news. Temporal data relating to the engagement of news, text properties and user properties were fed into the LSTM. The approach was tested on the datasets mentioned in Ma *et al*. (2016), attaining accuracy scores of 0.89 and 0.95 respectively. Kochkina *et al*. (2018) applied different methods based upon sequential models, namely LSTM, Linear and Tree Conditional Random Fields and Hawkes Process, for stance classification. The features derived from the sequential interaction of Twitter users were exploited so as to determine the stance taken by a tweet. LSTM emerged as the top performer. Furthermore, the authors proposed a "multi-task learning" strategy as a solution to classifying rumours. An LSTM layer was used to create a learning framework that was shared amongst the tasks, and several task layers. A per-event accuracy from 0.36 to 0.64 was recorded on the datasets.

Recent studies have capitalised on a combination of CNNs and RNNs alongside an LSTM layer in their models. A hybrid approach was proposed by Wang (2017) to detect fake news by using the LIAR dataset. The architecture adopts a CNN to encode text information (Song *et al*., 2019), and an LSTM layer for encoding the author's metadata. This hybrid model outperformed all the baselines, as well as a bi-LSTM, scoring 0.27 for accuracy. Ajao *et al*. (2018) carried out experiments on both a hybrid LSTM-CNN and a LSTM. The simplest LSTM model delivered the best performance with 0.82 accuracy. Furthermore, the authors have asserted that the model will probably yield an even higher accuracy on a larger dataset. Song *et al*. (2019) used the repost of sequence patterns in detecting false rumours found on social media. In particular, the authors used a CNN in the extraction of feature vectors associated with all posts, and these served as inputs for employing a "threshold strategy". The strategy was applied to the datasets mentioned in Ma *et al*. (2016), realising accuracies in excess of 0.90.

2.8 Meta-analysis of Fake News Detection Algorithms

The meta-analysis table presented on the following four pages was compiled during the review of relevant literature in this study. The analysis aided in informing the selection of methods used for

data pre-processing, feature extraction, and the metrics implemented during model performance appraisal. In addition, it influenced the adoption of an ensemble stategy, an elaboration of which is presented in the following chapter.

Table 2.2: Meta-analysis of Fake News Detection Algorithms

Title of Paper	Author / Year	Method	Approach	Performance Metrics	Number of Datasets Used
"Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News"	Rubin, V. L., Conroy, N., Chen, Y. & Cornwell, S., 2016.	Topic-based classification methodology with TF-IDF weighting scheme.	Machine learning.	f1 score = 087 precision = 0.90 recall = 0.84	1
"Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques"	Ahmed, H., Traore, I. & Saad, S., 2017.	N-grams.	Machine learning.	accuracy = 0.92	2
"Evaluating Machine Learning Algorithms for Fake News Detection"	Gilda, S., 2017.	Context-free grammars; TF-IDF.	Machine learning.	accuracy = 0.722 recall = 0.453 precision = 0.888 ROC = 0.883	1
"Automatic Detection of Fake News"	Pérez-Rosas, V., Kleinberg, B., Lefevre, A. & Mihalcea, R., 2017.	N-grams; Punctuation; Context-free grammars.	Machine learning.	accuracy = 0.78 f1 score = 0.79 precision = 0.75 recall = 0.84	2
"CSI: A Hybrid Deep Model for Fake News Detection"	Ruchansky, N., Seo, S. & Liu, Y., 2017	Textual, temporal, & user features.	Deep learning; Hybrid approach.	f1 score = 0.954 accuracy = 0.953	2
"Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection"	Wang, W. Y., 2017.	Surface-level linguistic patterns; Word2Vect.	Machine learning; Deep learning.	accuracy = 0.270	1
"Detecting Fake News with Machine Learning method"	Aphiwongsophon, S. & Chongstitvatana, P., 2018.	TF-IDF.	Machine learning; Deep learning.	f1 score = 0.998 accuracy = 0.999 precision = 0.998 recall = 0.999	1
"Determining fake statements made by public figures by means of artificial intelligence"	Granik, M., Mesyura, V. & Yarovyi, A., 2018.	TF-IDF.	Machine learning; Deep learning.	accuracy = 0.86	1
"FAKE STATEMENTS DETECTION WITH ENSEMBLE OF MACHINE LEARNING ALGORITHMS"	Granik, M. O. & Mesyura, V. I., 2018.	TF-IDF.	Machine learning; Deep learning; Ensemble learning.	accuracy = 0.88	1
"The Detection of Fake Messages using Machine Learning"	Looijenga, M. S., 2018.	Bag-of-Words.	Machine learning.	f1 score = 0.88, precision = 0.88 recall =0.88.	1
"A Deep Ensemble Framework for Fake News Detection and Classification"	Roy, A., Basak, K., Ekbal, A. & Bhattacharyya, P., 2018.	GloVe.	Ensemble learning; Deep learning.	f1 score = 0.43precision = 0.55 recall = 0.45	1

Title of Paper	Author / Year	Method	Approach	Performance Metrics	Number of Datasets Used
"Fake News Detection: A Deep Learning Approach"	Thota, A., Tilak, P., Ahluwalia, S. & Lohia., 2018.	Bag-of-Words; Tf-idf: GloVe; Word2Vect.	Deep learning.	accuracy = 0.942	1
"Automatic Online Fake News Detection Combining Content and Social Signals"	Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin,G., DiPierro, M. & de Alfaro, L., 2018.	Combination of social & content features.	Machine learning.	f1 score = 0.879 accuracy = 0.817 precision = 0.910 recall = 0.850	3
"TI-CNN: Convolutional Neural Networks for Fake News Detection"	Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z. & Yu, P. S., 2018.	Latent features - textual & visual.	Deep learning.	f1 score = 0.9210 precision = 0.9220 and recall = 0.9277	1
"Analysis of Classifiers for Fake News Detection"	Agarwal, V., Sultana, H. P., Malhotra, S. & Sarkar, A.,2019.	Bag-of-Words; N-grams; Tf-idf; Count Vectorizor.	Machine learning.	f1 score = 0.62 precision = 0.62 recall = 0.61	1
"Ensemble Learning Approach on Indonesian Fake News Classification"	Al-Ash, H. S., Putri, M. F., Mursanto, P. & Bustamam, A., 2019.	TF-IDF.	Ensemble learning.	f1 score = 0.95 precision = 0.965 recall = 0.93	1
"A Topic-Agnostic Approach for Identifying Fake News Pages"	Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E. & Freire, J., 2019.	Linguistic & web-markup features.	Machine learning; Topic-agnostic approach.	accuracy = 0.86	3
"Multi-Perspective Ensemble for Hyper-Partisan News Detection"	Colgan, W. & Kakkar, K., 2019.	Word & sentence embedding.	Ensemble learning.	accuracy = 0.93	1
"Multiclass Fake News Detection using Ensemble Machine Learning"	Kaliyar, R. K., Goswami, A. & Narang, P., 2019.	TF-IDF; CoSine similarity; Word overlap; Polarity; Refuting; Hand-selected features.	Machine learning; Ensemble learning.	accuracy = 0.86	1
"A Benchmark Study on Machine Learning Methods for Fake News Detection"	Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A. & Afroz, S., 2019.	N-grams; TF-IDF; Word embedding; Empath Tool; Sentiment Intensity Analyzer	Machine learning; Deep learning.	f1 score = 0.95recall = 0.95 precision = 0.95 accuracy = 0.95	3
"Stacking-based Ensemble Learning on Low Dimensional Features for Fake News Detecting"	Li, S., Ma, K., Niu, X., Wang, Y., Ji, K., Yu, Z. & Chen, Z., 2019.	Bag-of-Words; Low dimensional feature extraction; TF-IDF; Two-stage stacking fusion.	Ensemble learning.	accuracy 0.98	1
"A Location Independent Machine Learning Approach for Early Fake News Detection"	Liu, H., 2019.	Bag-of-Words; GloVe.	Ensemble learning.	f1 score =0.942 accuracy = 0.961 precision = 0.980 recall = 0.907	1
"Identification of Fake News Using Machine Learning"	Mandical, R. R., Mamatha, N., Shivakumar, N., Monica, R. & Krishna, A., 2019.	TF-IDF; Bag-of-Words; Tokenizers.	Deep learning; Machinelearning.	accuracy = 1.00	8
"Fake News Detection Using Ensemble Machine Learning"	Mohale, P. & Leung, W. S., 2019.	TF-IDF.	Machine learning; Ensemble learning.	f1 score = 0.82 precision = 0.80 recall = 0.76	1

Title of Paper	Author / Year	Method	Approach	Performance Metrics	Number of Datasets Used
"Fake Data Analysis and Detection Using Ensembled Hybrid Algorithm"	Reddy, P. B. P., Reddy, M. P. K., Reddy G. V. M., & Mehata, K., 2019.	Bag-of-Words.	Machine learning; Ensemble learning; Hybrid approach.	f1 score = 0.9193 precision = 0.909 recall = 0.9298	1
"An ensemble approach for spam detection in Arabic opinion texts"	Saeed, R. M., Rady, S. & Gharib, T. F., 2019.	N-grams; Negation handling.	Ensemble learning.	accuracy = 0.9998 f1 score = 0.9998 specificity = 0.9997 recall =0.9998 precision = 0.9999	2
"A Unique Approach for Detection of Fake News using Machine Learning"	Shukla, Y., Yadav, N. & Hari, A., 2019.	N-grams; TF-IDF.	Machine learning.	f1 score = 0.718	1
"Ensemble Learning Approach For Clickbait Detection Using Article Headline Features"	Sisodia, D. S., 2019.	19 manually selected features from article headlines.	Ensemble learning; Machine learning.	f1 score = 0.9116 accuracy = 0.9116 precision = 0.9116 recall = 0.9116	5
"Performance Comparison of Machine Learning Classifiers for Fake News Detection"	Smitha, N. & Bharath, R., 2019.	Count Vectorizer; Word embedding; TF-IDF.	Deep learning; Machine learning	precision = 0.93 accuracy = 0.94 recall = 0.93 f1 score= 0.93	1
"Fake News Detection Using Machine Learning"	Waikhom, L. & Goswami, R. S., 2019.	N-grams; TF-IDF.	Machine learning.	f1 score = 0.70 precision = 0.70 recall = 0.70	1
"Fake News Detection: An Ensemble Learning Approach"	Agarwal, A. and Dixit, A., 2020.	Bag-of-Words;; TF-IDF, N-grams; Word2Vect; POS-Tagging.	Machine learning; Deep learning; Ensemble learning	f1 score = 0.965 recall = 0.97 precision = 0.96 accuracy = 0.97	2
"Fake News Detection Using Machine Learning Ensemble Methods"	Ahmad, I., Yousaf, M., Yousaf, S.& Ahmad, M. O., 2020.	Linguistic Inquiry; Word Count.	Machine learning; Deep learning; Ensemble learning	f1 score = 0.99 precision = 0.99 recall = 1	4
"Fake Reviews Detection through Ensemble Learning"	Gutierrez-Espinoza, L., F., Namin, A. S., Jones, K. S. & Sears, D. R., 2020.	Doc2Vec.	Ensemble learning	accuracy = 0.773	1
"Automating Fake News Detection System using Multi-level Voting Model"	Kaur, S., Kumar, P. & Kumaraguru, P., 2020.	TF-IDF; Hashing Vectorizer; Count Vectorizer.	Ensemble learning	f1 score = 0.988accuracy = 0.989 precision = 0.991 specificity = 0.982 recall = 0.987	3
"GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media"	Lu, Y. J. & Li, C. T., 2020.	Word embedding.	Deep learning	f1 score 0.759 accuracy = 0.908 precision = 0.759 recall = 0.763	2
"A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers"	Mahabub, A., 2020.	TF-IDF.	Machine learning; Deep learning; Ensemble learning	f1 score = 0.95 recall = 0.95 precision = 0.95	1

Title of Paper	Author / Year	Method	Approach	Performance Metrics	Number of Datasets Used
"A Language-Based Approach to Fake News Detection Through Interpretable Features and BRNN"	Qiao, Y., Wiechmann, D. & Kerz, E., 2020.	Automated text analysis - CoCoGen.	Deep learning; Language-based approach	accuracy = 0.993 precision = 0.993 recall = 0.993	2
"Profiling Fake News Spreaders on Twitter Notebook for PAN at CLEF 2020"	Saeed, U., Fahim, H. & Shirazi, F., 2020,	Emphasis on converting all stylistic information into unique tags.	Machine learning; Deep learning	accuracy = 0.70	1
"An Ensemble Technique to Detect Fabricated News Article Using Machine Learning and Natural Language Processing Techniques"	Sangamnerkar, S., Srinivasan, R., Christhuraj, M. & Sukumaran, R., 2020.	Doc2Vec.	Ensemble learning	f1 score = 0.8808 accuracy = 0.8808 precision = 0.8809 recall = 0.8808	1
"Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation"	Shu, K., Mahudeswaran, D., Wang, S. & Liu, H., 2020.	Features from structural, temporal, & linguistic perspectives.	Deep learning; Hierarchical propagation network	f1 score = 0.862 accuracy = 0.861 precision = 0.854 recall = 0.869	1
"Towards automatically filtering fake news in Portuguese"	Silva, R. M., Santos, R. L., Almeida,T. A. & Pardo,T. A., 2020.	Bag-of-Words; Word2Vec; FastText.	Machine learning	false +ve rate = 0.036 f1 score = 0.971 precision = 0.964 recall = 0.978	1
"Fake News Detection with Different Models"	Vijayaraghavan, S., Wang, Y., Guo, Z., Voong, J., Xu, W., Nasseri, A., Cai, J., Li, L., Vuong, K. & Wadhwa E., 2020.	TF-IDF; Count Vectorizer; Word2Vec.	Machine learning; Deep learning	accuracy = 0.9488	1
"SAFE: Similarity-Aware Multi-modal Fake News Detection"	Zhou, X., Wu, J. & Zafarani, R., 2020.	Muli-modal feature extraction.	Deep learning	f1 score = 0.896accuracy = 0.874 precision = 0.889 recall = 0.903	2
"Linguistic feature-based learning model for fake news detection and classification"	Choudhary, A. & Arora, A., 2021.	Linguistic model - syntactic, sentimental, grammatical, & readability.	Deep learning.	accuracy = 0.72	2

2.9 Chapter Summary

This chapter has comprehensively reviewed relevant literature so as to accomplish the first objective of this study and to serve as a foundation and motivation for the current study. This chapter comprised seven sections that provided detailed descriptions of pertinent topics related directly to the current study. Emphasis was placed on the different strategies and methods proposed in the literature by recognising their strengths and drawbacks. The detection of fake news is a

complex problem. Although numerous solutions have been proposed, the problem continues to grow. The next chapter expounds the methodology that was applied to achieve the primary aim of the study, which is to apply an advanced ensemble approach to enhance the detection of fake news.

CHAPTER THREE: RESEARCH METHODLOGY

3.1 Introduction

This chapter explains the methodological process employed in this study in order to achieve the overall research aim and specifically Objective 2 as outlined in Chapter 1. Chapter 3 is structured as four sections. Firstly, design research is discussed. This is followed by an explanation of the proposed methodological framework selected, and the cross-validation strategy utilised in this study. Finally, the algorithms that are used by the research study's proposed Blending Ensemble Model 1 and Blending Ensemble Model 2 are documented.

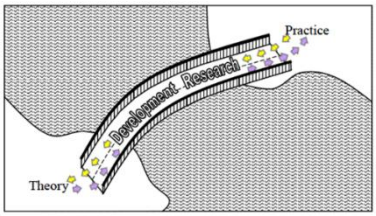
3.2 Design Research

Of late, a growing interest in "design research" and design from both academia and industry has been observed. "Design research" has become common parlance in the discipline of design, and is more frequently used in describing a myriad of methods, perspectives, approaches and philosophies. The term "design research", amalgamates the areas of design, practice and research, bringing about a union "roughly equivalent to the investigation of knowledge through purposeful design" (Faste & Faste, 2012). Several authors have expressed the opinion that design research includes both "the study of design and the process of knowledge production", which arises through the action of design (Fallman, 2007; Koskinen *et al*., 2011).

Irrespective of the domain, there are two significant features in research relating to design and development, namely: (1) the research culminates in the creation of an artifact; and (2) the practice is based on research (Ellis & Levy, 2010). Artifacts comprise the creation of a fresh product, process, or tool (Richey & Klein, 2007). An artifact may also include less evident results, for example, the development of new design and development models, new theories, new processes and methods, and formerly untested models, methods or tools for solving a problem in a completely different context (Hevner *et al*., 2004).

Research of this nature can be envisaged as executing a "bridging function" during the research cycle. It commences with conceptualisation of the problem and concludes with the appraisal of the artifact(s) regarding their efficacy in solving the problem. The primary focus is on using the central concept to build the bridging artifact, as exemplified by the diagram below.

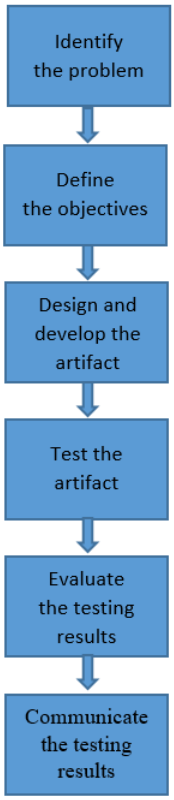
Figure 3.1: Design framework (Ellis & Levy, 2010, p. 109).



3.2.1 Design and Development Activities

There exists a variance in terms of the names, numbers and key milestones in the literature relating to design and development. Ellis and Levy (2010) developed a six-phase framework based primarily on methodology from systems development that incorpotated concepts from Peffers *et al*. (2007). The framework that has been adopted for this study is depicted in the flow chart below. Each phase will be discussed separately thereafter.

Figure 3.2: Design and development framework (Ellis & Levy, 2010).



3.2.1.1 Identify the problem

A well-articulated problem statement is the crucial starting point. This activity defines the research problem clearly in order to identify a probable solution. It is imperative that the problem be articulated conceptually to ensure its full complexity is captured. Knowledge regarding the problem and the importance of finding a solution to the problem is pivotal. The problem and proposed solution with respect to this study are summarised below.

There is a scarcity of effective tools available to enlighten the public regarding the authenticity of news (Khan *et al*., 2019). As such, the problem of false news propogation continues to burgeon to the extent that manual efforts are no longer able to curtail it. Consequently, the automated detection of fake news has become critical. Classical machine learning has been applied in detecting misinformation with varying degrees of success yet numerous advanced models have not been applied. Current studies are highlighting the potential of ensemble learning (Gutierrez-Espinoza *et al*., 2020), and this research intends to explore this potential with the aim of addressing the gap.

3.2.1.2 Define the objectives for a solution

Clearly defined objectives of any research endeavour should underlie the study. These objectives must be directly linked to the problem but should not have documented solutions. The objectives are derived from the definition of the problem and the knowledge related to the circumstances concerning the problem. Furthermore, it is essential to be transparent about what is feasible or probable in respect of every objective. An account of how the envisaged artifact will aid in solving the problem is integral.

The resources necessary for this task are an in-depth understanding of the problem and the probable solution. The objectives of this study are as follows:

- To comprehensively research relevant publications based on the detection of fake news with the objective of identifying methods and approaches to devising a probable solution.

- To develop an ensemble approach for detecting fake news by combining machine learning with natural language processing.

- To experimentally evaluate the developed model against other machine learning algorithms making use of well-known statistical evaluation metrics for comparison.

3.2.1.3 Designing and developing the artifact

An artifact is created during this activity phase. The design and development process should be anchored in current, relevant literature. Generally, this process involves the following steps: (1) constructing a conceptual framework; (2) designing the system architecture; and (3) creating a prototype for testing and evaluation. In such a process, both the architecture and functionality of the artifact are considered. In this study, machine learning, advanced ensemble learning, and natural language processing strategies, will all be harnessed in the development of the artifact.

3.2.1.4 Test the artifact

It is essential to display how an artifact that has been created fulfills the requirements and functionalities proposed in the design and development stage. Another vital component of testing and appraisal is the suitability of the developed artifact to the problem setting. The prototype should indeed be applicable and appropriate to the proposed context and should also demonstrate feasible results in tackling the problem. The methods employed during testing and evaluation will vary and depend on the characteristics of or type of artifact and the availability of resources. In this study, the models will be tested on two widely used and benchmarked fake news datasets.

3.2.1.5 Evaluate the testing results

This activity necessitates measuring and observing how effective the solution is in solving the problem. Knowledge of the relevant metrics and analysis techniques plays a principal role during this phase as the researcher will be guided by the evaluation process in order to determine whether a return to the third activity to make enhancements to the functionalities pertaining to the artifact is required. This process may need to be repeated a number of times. A combination of metrics, for example recall, roc, accuracy, auc, f1-score, precision, and confusion matrix will be applied in appraising the execution of the models.

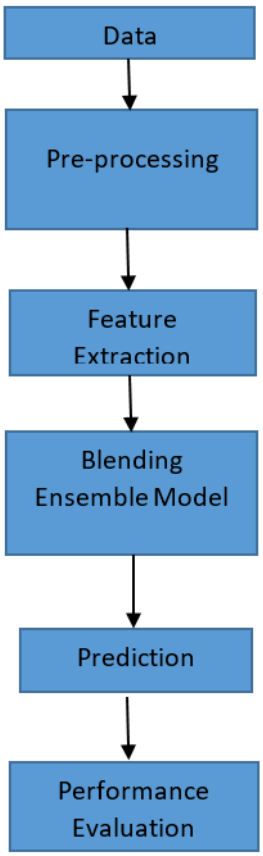
3.2.1.6 Communicate results and conclusions

Complete and clear reporting of the conclusions and outcomes of a research study represent the impact and significance of the study for other researchers in the field. It is therefore imperative to document and disseminate such results properly. The research problem, the artifact design, and its utility should be shared with interested parties. In this study, the research problem, solution and outcomes will ultimately be published in scholarly journals.

3.3 Proposed Methodological Framework

The framework for the selected methodology most applicable to the proposed model, as shown in the flow chart below, extends the existing literature by integrating blending into current ensemble learning practices. News reports from numerous sources are classified as “true” or “fake” by employing different sets of features. A blending ensemble consisting of term frequency, inverted document frequency, n-grams, and word embeddings, is utilised in the methodology.

Figure 3.3: Proposed Methodology Model



3.3.1 Datasets

The ascertainment of news as "fake news" can be "a very challenging and time-consuming task" (Agarwal *et al*., 2019). Therefore, datasets that are currently available have been selected for this study. An immense challenge encountered in identifying bogus news is the quality and availability of appropriate datasets (Oshikawa *et al*., 2018). Furthermore, being able to use a collection of reports without hinderance can be problematical due to existing copyrights (Gilda, 2017). The Liar (Wang, 2017) and ISOT (Ahmed *et al*., 2018) datasets have been selected for this project. Both are open source and therefore readily available online. The datasets contain news in a mixture of truthful and fake reports from a variety of categories.

3.3.1.1 Liar Dataset

The dataset is easily accessible with no restrictions and has effectively been employed in prior studies (Mohale & Leung, 2019). It incorporates 12836 brief, labelled reports sourced from politifact.com. Six rating labels are applied in ranking the reliability of the reports: (1) "pants-fire"; (2) "false"; (3) "barely-true"; (4) "half-true"; (5) "mostly-true": and (6) "true". The emphasis is on the classification of news. To apply binary classification, the six rating labels are transformed into two. "Pants-fire", "false", and "barely-true" are collectively classified as "fake", and "half-true", "mostly- true", and "true" are all labelled as being "true". The Liar dataset principally comprises of political statements involving US democrats and republicans, and a substantial number of posts from social networks (Khan *et al*., 2019).

3.3.1.2 ISOT Dataset

The ISOT dataset encompasses both fake and truthful articles derived from numerous domains (Ahmad *et al*., 2020). The truthful reports originate mostly from reuters.com, which is a recognised news site on the web. The fake news articles were acquired from a broad range of sources, chiefly from online sites identified by politifact.com, which specialises in checking the validity of claims. The dataset has a sum total of 44,898 reports; of which 21,417 are truthful reports, and 23,481 bogus reports. Individual data points comprise the date, title, text, and subject. The text represents the news report, with the subjects including news from the Middle East, the US, the government, politics news, left news and world news.

3.3.2 Pre-processing

"Raw" texts containing news require pre-processing prior to being consumed by the models. This task is facilitated by applying natural language processing techniques, which also contribute towards improving overall accuracy. The sequencing of operations executed are:

- Cleansing the data – filter out superfluous data which is not needed during analysis such as duplicate data.

- Check for absent data that could have an unfavourable impact on the final outcome. For example, an incomplete article.

- Transform all text into lowercase for consistency during processing.

- Eliminate all punctuation as these are not required during the analysis.

- Removing stopwords present in the text. Stopwords have no importance during the processing of natural language because they do not provide additional semantical meaning. Examples include, "a", "the", "are", "is" and etc.

- Stemming (lemmatization) of the data to transform words into their original forms. For example, the three grammatical forms of "Dancing", "Dance", and "Dancer" will be dimished to the original noun, which is "dance". This reduces the number of classes in the dataset and enables speedier and more efficient classification (Ahmed *et al*., 2017). Stemming will be implemented using the Porter algorithm on account of its accuracy.

3.3.3 Feature Extraction

The design of suitable features plays a significant part in the performance of a learning model. Extracting the most pertinent words as features "can be extremely useful" (Onan *et al*., 2016). Term Frequency (TF), n-grams, Term Frequency Inverted Document Frequency (TF-IDF), and word embeddings, will be utilised during the feature extraction process.

TF is used to determine likeness among documents using counts of the pertinent words. Vectors of equal dimensions that contain the word count are used to represent each document. Each vector

is normalised giving its elements a sum of one. All word counts are expressed as a probability of these words being in the documents. If a word is present it will be assigned one, otherwise, it will be given the value zero. A document is therefore denoted by word clusters. TF is expressed by the formula (Smitha & Bharath, 2020):

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad (3.1)$$

where n represents the total archives term t appears in a word.

The metric TF-IDF is commonly used in information retrieval and for processing natural language. It measures how important a term is in a document. A term's significance escalates with the number of its occurrences within the document. This is counterbalanced by its frequency in the corpus. A key characteristic of Inverted Document Frequency (IDF) is that it "weighs down the term frequency while scaling up the rare ones". For example, the word "the" appears frequently in texts and will be dominant in the frequency count if only TF is applied. By using IDF, the impact of these terms is downscaled. Both IDF and TF are used to compute TF-IDF (Smitha & Bharath, 2020):

$$IDF = \log(N/n) \quad (3.2)$$

$$TF\text{-}IDF = (IDF * TF) \quad (3.3)$$

where N represents the total archives and n denotes the total archives a term t appears in a word.

N-grams are adjacent sequences of objects that have dimension "n" and can consist of bytes, characters or syllables, and words. N-grams based on words and characters are mostly applied for text classification (Ahmed *et al*., 2017). Word-based n-grams are utilised to represent the context of a document and to generate features and can be useful in the classification of a document as real or fake. This approach has been used successfully with bigrams and unigrams (Khan *et al*., 2019).

Word embeddings represent words in the form of a dense vector. These embeddings are trained on the input data or created from pre-trained embeddings. Pre-trained word embeddings will be used in this study. These embeddings have been initialised using the one hundred dimension pre-trained GloVe embedding. "GloVe is an unsupervised learning algorithm" utilised for finding the vector

equivalents of words and was trained using a billion words and a vocabulary comprising four hundred thousand words (Khan *et al*., 2019). GloVe embedding is used since it is not dependent on the datasets.

3.3.3 Machine Learning Algorithm

The algorithm will carry out the following processes:

- Divide the dataset into testing and training datasets.
  The model will be trained by consuming the training data while the testing data will be used for evaluating the model.

- Create the base models (sub-models)

- Build the blending ensemble model by conflating the selected sub-models.
  The predictive power of the individual sub-models will be harnessed, and then fused to build the ensemble model.

3.3.4 Prediction

The outcome of the prediction made using the proposed model will be presented and the result will reflect whether the news article is either true (real) or fake.

3.3.5 Performance Evaluation

Different metrics have been used to evaluate the performance of models. The main objective of a machine learning model is to generalise or perform well on new data and performance metrics assist in quantifying this performance (Chauhan 2020). Consequently, the model performance can be improved by tuning hyper parameter(s) or tweaking the features gathered from the input dataset. Performance metrics are thus included in all machine learning pipelines. They are indicators of the progress that has been made, which is represented by a number. Metrics are necessary to judge performance, and are applied to measure and monitor performance during both testing and training.

The concept of creating machine learning models is based on a "constructive feedback principle". The model is built, metrics are used to provide feedback, and improvements are made, until the acceptable level of accuracy is ultimately achieved (Srivastava, 2019). Without implementing an appropriate evaluation of the machine learning model using a variety of metrics, and thereby relying on accuracy alone, it can become problematic when the model encounters unseen data, resulting in poor predictions (Chauhan, 2020). A combination of metrics is therefore used in this study. In particular, the metrics utilised include roc auc, accuracy, auc, precision, recall, confusion matrix and f1-score. These metrics will be explained in the sections that follow.

3.3.5.1 Roc Auc

This metric calculates the area underneath the "receiver operating characteristic" (ROC) curve by using the scores of the prediction. The roc auc can be computed by using the Python roc_auc_score() function. The true outcomes and the probabilities that have been predicted are used and the score computed ranges from 0 (no skill) to 1 (perfect skill).

This can be taken as a summary of skill associated with the model. It measures the total two-dimensional area beneath the curve using the interval (0,0) to (1,1), and delivers an aggregate score for the performance for all classification thresholds. The equation for computing the area beneath the curve is:

$$\mathrm{AUC(T)} = \sum \mathrm{k\overline{P}\ kD(T)\ \Delta PkFA(T)} \qquad (3.4)$$

where $\Delta \mathrm{PkFA(T)} = \mathrm{Pk+1FA(T)} - \mathrm{PkFA(T)}$ and $\overline{\mathrm{P}}\mathrm{kD(T)} = \mathrm{P(k+1)D(T)} + \mathrm{PkD(T)2}$ (Sadiq, 2018).

3.3.5.2 Auc

This metric represents the area below the precision recall curve. It is computed by using the recall and precision scores determined for every threshold. The auc can be computed using the Python auc() function. It summarises an approximation or the integral of the region beneath the curve, which can be approximated using (Zhubarb, 2018):

$$\sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{P(k)\ \Delta r(k)} \qquad (3.5)$$

An auc score of 0 indicates that a model's predictions are 100% incorrect. A perfect classifier that does not make any incorrect predictions will have an auc score of 1.

3.3.5.3 Accuracy

The accuracy metric is the most commonly used. Correct predictions made by the model are indicated as percentages. Accuracy is calculated by using the equation below (Ahmad *et al*., 2020):

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TN+TP+FN+FP}} \tag{3.6}$$

where TP denotes the true positives, FP the false positives, TN the true negatives, and FN the false negatives.

Generally, a high accuracy value represents a favourable model and vice versa, however, if a true news article was predicted as false or the other way around, this could have adverse consequences and subsequently issues of trust begin to become a concern. Other metrics are therefore applied to account for classification errors.

3.3.5.4 Recall

The capacity of a classifier to discover every positive instance is called recall. It can be determined for every class as the ratio of the TP to the summation of the TP and FN. Recall denotes the total number of the positive classifications from the true class. It will represent the total number of news reports forecasted as true out of the overall number of true reports (Agarwal *et al*., 2019).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3.7}$$

3.3.5.5 Precision

Precision represents the capacity of a classifier to avoid mislabelling a case positive, when it is really the opposite. It is defined for each class as the ratio of TP to the summation of TP and the FP. The precision value will represent the number of news reports indicated as true out of the all true (predicted positively) reports (Agarwal *et al*., 2019):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3.8)$$

3.3.5.6 F1 Score

This is the harmonic mean between recall and precision. It is a metric applied during performance rating. The maximum (best) score is 1 and the minimum (worst) is 0. Usually this score is less than the accuracy score since it embeds precision and recall in the calculation.

F1-scores represent the trade-off between recall and precision. Hence, it takes both FN and FP into consideration. The score can be computed by applying the following formula (Agarwal *et al*., 2019):

$$\text{F1} - \text{score} = 2\frac{\text{Recall x Precision}}{\text{Recall+ Precision}} \quad (3.9)$$

3.3.5.7 Confusion Matrix

This is a matrix with dimension (N x N) and is used in assessing the performance by a model during classification, where the number of target classes is N. It is a table (matrix) that aids in giving a better understanding of a classification model's performance. The classifier's accuracy is visualised by associating the predicted and actual classes (Shukla *et al*., 2019). The table comprises the four different combinations of the actual and predicted values. The actual target values are compared to those predicted by the machine learning model. The predicted values of the target variable are represented by the rows. The outcomes of the confusion matrix are shown below.

Table 3.1: Confusion Matrix.

		FALSE	TRUE	
Actual	FALSE	True Negative	False Positive	FALSE
	TRUE	False Negative	True Positive	TRUE
		Predicted		

3.4 Cross-Validation

This is an established technique that is widely used in applied machine learning to test the efficacy of the models. It is a resampling method used in model evaluation when the data is limited. During cross-validation, a portion/sample of the data is reserved and is not used in training the model. At a later stage, it will be used for validation or testing. To measure the

performance of a model it is essential to carry out testing on some data that has not been seen before. The results obtained from this process will aid in determining whether a model is over-fitting, under-fitting or well-generalised (Medium, 2018). The cross-validation methods adopted in this study include the train-test split approach and k-fold cross-validation.

3.4.1 Train-test Split

Here the data is split randomly into testing and training sets. The training set is used during model training, and the testing set when testing is performed. As a general rule the dataset is split into an 80:20 or 70:30 train-test ratio, however there is no set prescription. It must be noted though that there is a likelihood of significant bias being introduced if the data is of a limited nature, since some of the information related to the data that was not used during training may be overlooked. Nevertheless, for a large dataset with test and train samples having a similar distribution, this approach is more appropriate and more acceptable (Brownlee, 2020).

3.4.2 K-fold Cross-validation

This process ensures that each data point in the original set has the opportunity of featuring in the training and testing sets. Consequently, there is less bias when compared to other approaches. The procedure makes use of a parameter named k, which is the number of portions a dataset will be divided into and as such is called k-fold cross-validation. The method randomly distributes the data points into k-folds of almost equivalent size. The first fold will eventually be used in validating the model, and the residual (k – 1) folds is used to fit the model (Brownlee, 2018). This approach is well-suited to working with limited data.

The method is summarised in the steps described below:

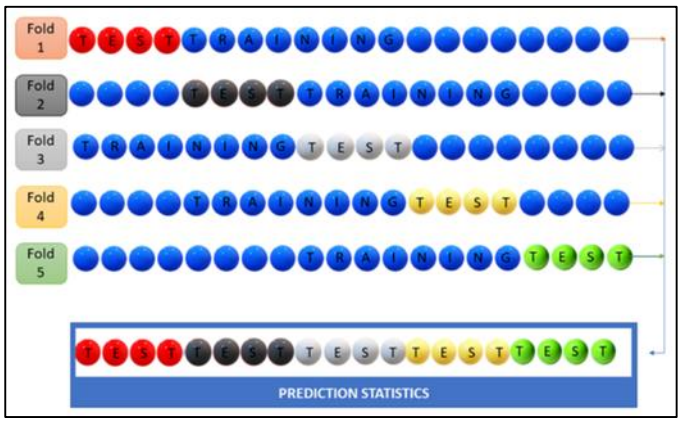
> Step 1. Split the entire dataset randomly into k-folds. Usually, k=5 or k=10 is chosen dependent on the data size but the value of k should not be too high or too low. A high value of k may result in a less biased model, but a large variance could lead to over-fit. The lower value of k is analogous to the train-test split approach.

Step 2. Fit the model using the (k-1) folds, and validate the model using the kth fold. Record the scores and errors.

Step 3. Repeat this process until every k-fold has served as a test set. Then, compute the average of the recorded scores, which will result in the performance metric of the model.

Figure 3.4 below illustrates the process graphically.

Figure 3.4: K-fold Cross-validation (Data Science Central, 2019).



3.5 Algorithms

The subsequent learning algorithms in combination with the methodology that has been proposed are used to gauge how the classifiers perform in the detection of news that is fake.

3.5.1 Blending Ensemble Model 1

Blending is a variant of stacking. Despite the fact that they are closely related there is one significant difference, being that the predictions based on a holdout validation dataset is used to fit the meta-model. The model learns to blend predictions derived from the base models (sub-models) to eventually reach the final prediction. The Blending Ensemble Algorithm is given below:

1. Divide dataset.
    - Create training and testing sets.
    - Split the training set into hold-out and train-sets.

2. Build sub-models.

3. Train Blending Ensemble.
   - Repeat.
     Use training set to fit sub-model.
     Predict a sub-model using hold-out set.
     Record predictions for blending input.
     Until end of sub-models.
   - Transform stored predictions into 2D matrix.
   - Build blending model.
   - Use stored predictions (2D matrix) to fit blending model.

4. Predict on Blending Ensemble.
   - Repeat.
     Predict on sub-model using test set.
     Store prediction.
     Until end of sub-models.
   - Transform stored predictions into 2D matrix.
   - Predict on blending model using stored predictions (2D matrix).

5. Evaluate the predictions.

The description of the algorithms included in Blending Ensemble Model 1 are presented next.

3.5.1.1 Logistic Regression

Logistic regression is chosen since the classification of the text gives a binary output (true/fake or true/false or 0/1). The hypothesis function is expressed as follows:

$$h_\theta(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}. \quad (3.10)$$

The output is transformed into a probability by applying a sigmoid function. The objective is to optimise the probability through minimisation of the cost function displayed below (Ahmad *et al*., 2020).

$$Cost\ (h_\theta(x), y) = \begin{cases} log(h_\theta(x)),\ y = 1, \\ -log(1 - h_\theta(x)),\ y = 0. \end{cases} \quad (3.11)$$

Consequently, the logistic curve is constrained to values ranging between 0 and 1 by employing a sigmoid function.

3.5.1.2 Support Vector Machine

Support Vector Machines (SVMs) create a hyperplane for isolating and grouping of features. On both sides of the hyperplane support vectors are used to calculate the ideal hyperplane with the vectors maximising the space amongst each other. The bigger the vector space on all sides of the hyperplane culminates in a more precise decision boundary among the class features. (Vijayaraghavan *et al*., 2020).

The data points are categorised into individual groupings depending on where they are located in the hyperplane. The aim is to maximise the existing space between the data points and hyperplane. The margin is maximised by a loss function. The hyperplane equation is:

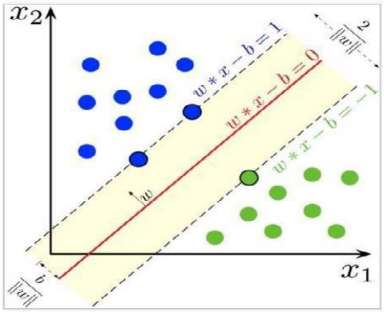
$$w^x + b = 0 \quad (3.12)$$

where b is the bias, and w the weight vector. The Loss function is formulated as:

$$L(w) = \sum_1 max\ (0,1 - y_i[w^t x_i + |b|]) + \lambda||w||^2 \quad (3.13)$$

Errors are computed by the first term. The regularisation function is denoted by the second term and is applied to sidestep overfitting (Agarwal & Dixit, 2020). Figure 3.5 below depicts the "maximum-margin hyperplane" and margins for two classes.

Figure 3.5: SVM Analysis (Garg *et al*. 2020, p.7).



3.5.1.3 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) algorithms are usually used for classification problems. They calculate a summary of statistics, such as standard deviation and average, linked to the features used for input by class label. These statistics reveal what has been learnt by the model from consuming the training data.

Predictions are based on probability estimates of a new case matching a class label. The class that has the greatest probability is allocated to this case. LDA can be perceived as a straight forward implementation of Bayer's Theorem directed at classification.

The process is summarised as follows (Mehta, 2020):

(a) Compute the "between-class variance" or separability among the individual classes. The formula is:

$$S_b = \sum_{i=1}^{g} N_i(\bar{x_i} - \bar{x})(\bar{x_i} - \bar{x})^T \quad (3.14)$$

(b) Compute the "within-class variance" applying the formula:

$$S_w = \sum_{i=1}^{g}(N_i - 1)S_i = \sum_{i=1}^{g}\sum_{j=1}^{N_i}(x_{i,j} - \bar{x_i})(x_{i,j} - \bar{x_i})^T \quad (3.15)$$

(c) Create a lower-dimensional space to minimise S_w and to maximise S_b. The projection of the lower-dimensional space is expressed as (Fisher's criterion):

$$P_{lda} = \arg\max_{P} \frac{|P^T S_b P|}{|P^T S_{wP}|} \quad (3.16)$$

LDA makes the assumption that numeric inputs have a normal distribution with equivalent variance or spread. Otherwise, it may be required to normalise or convert data prior to modelling. It supports dual-class problems and multi-class classification without any modifications.

3.5.1.4 Stochastic Gradient Descent

This algorithm uses iteration in the optimisation of an objective function through suitable smoothness attributes such as differentiables or sub-differentiables. The strategy uses randomly chosen samples in gauging the gradients. Hence, stochastic gradient descent "can be regarded as a stochastic approximation of gradient descent optimization" (Shukla *et al*., 2019).

The gradient is essentially the incline or slope of the function. Moreover, it can be viewed as the amount "of change of a parameter with the quantity of change in another parameter" (Garg *et al*., 2020). The steeper the slope the higher the gradient. Gradient descent works iteratively to determine the values of a function's parameters so as to "minimize the function value with the maximum quantity". Hence, the goal is to find optimal values for the parameter in order to acquire the smallest cost function value (Hansrajh *et al*., 2021).

These details can be formulated for classification as, for a training set $(x_1, y_1) \ldots (x_n, y_n)$, where $x_i \in \mathcal{R}^m$ and $(y_i \in -1,1)$, with the aim being to "learn" a linear function $f(x) = w^t + b$ with parameters $w \in \mathcal{R}^m$ and intercept $b \in \mathcal{R}$. Binary classification predictions are carried out by examining the sign of f(x). The model parameters can be determined by minimising the regularised error in training, which is formulated below:

$$E(w, b) = \frac{1}{n}\sum_{i=1}^{n} L\big((y_i, f(x_{i)})\big) + \alpha R(w) \quad (3.17)$$

The loss function is L. R denotes a regularisation term utilised to penalise the model's complexity; $\alpha>0$ is the hyper parameter for controlling the regularisation strength (Scikit-learn, 2021).

3.5.1.5 K-Nearest Neighbour

K-nearest neighbour algorithm is mostly used for regression and classification. Predictions are made by taking the greatest historically similar examples when considering new data. The algorithm assumes that things that are similar exist nearby. That is, related data points would be close to one another. To make a prediction for the input data the k most similar cases are picked by applying a "distance measure". The Minkowski distance is given by:

$$(\sum_{1}^{n}|x_i - y_i|^p)^{\frac{1}{p}} \quad (3.18)$$

where the Manhattan distance is given when p = 1, and the Euclidean distance when p = 2 (Agarwal & Dixit, 2020).

3.5.1.6 Ridge Regression

The regression technique provides the basis for a Ridge Classifier. During binary classification the target variable is transformed to -1 or +1 depending on its class, and for multi-class data the prediction with the largest value determines the target class.

Linear regression and ridge regression are virtually comparable. The difference relates to a limited amount of bias being intrdoced during ridge regression resulting in a significant reduction in the variance. So, by commencing initially with a poor fit, improved predictions are possible eventually. The additional bias is termed "the ridge regression penalty". It is calculated by the product of the squared weight linked to individual features and lambda.

The foisting of a penalty based on the coefficient's size assists in overcoming some challenges of Ordinary Least Squares. The ridge coefficients minimise the penalised residual sum of squares:

$$\min_{w}||Xw - y||_2^2 + \alpha||w||_2^2 \quad (3.19)$$

$\alpha \geq 0$ represents a complexity parameter to control shrinkage. The larger the magnitude for α, the greater the shrinkage. Thus the coefficients are more robust towards collinearity (Scikit-learn, 2021).

3.5.2 Blending Ensemble Model 2

Blending Ensemble Model 2 comprises ensemble models that include a mix of bagging, boosting and stacking machine learning models to ensure diversity. The description of the algorithms included in Blending Ensemble Model 2 are presented next.

3.5.2.1 Voting Ensemble

Voting ensembles are usually used for classification problems where the final outcome is determined by the "majority vote" for a particular class (Granik & Mesyura, 2018). Every base model votes (predicts) once for the target class as determined by the model when a prediction is made for a statement. All the votes from the base models are tallied. The winning class is the class that has the most votes, and becomes the classification result. In the eventuality of a tie, the result of the classification is randomly selected from the classes with the highest vote count. Voting ensembles are simpler to implement than boosting and bagging algorithms (Ahmad *et al*., 2020). However, a voting ensemble is a combination of several non-dependent models that provide classification outcomes that form part of the overall prediction by implementing majority voting.

3.5.2.2 Random Forest

This algorithm, as mentioned before is an ensemble consisting of classification trees. Several decision tree classifiers are fit on numerous subsamples based on the dataset. Tree building uses random variable selection. Individual trees are fully grown to achieve low bias trees. Bootstrap aggregation (bagging) is used for uniting unstable learners. This, together with random selection of variables, produces a low correlation of trees. "The algorithm yields an ensemble that can achieve both low bias and low variance (from averaging over a large ensemble of low bias, high-variance but low correlation trees)" (Shukla *et al*., 2019). An RF classifier merges the predictions of numerous trees, which were trained individually. The classification accuracy can be enhanced by boosting the number of trees used. Furthermore, it can effectively deal with big variable inputs, and balancing of errors present in unbalanced datasets. RF gives greater performance than an individual decision tree in terms of classification (Al-Ash *et al*., 2019).

For a training set consisting of input samples $X = x_1, x_2, ..., x_n$ and corresponding output samples $Y = y_1, y_2, ..., y_n$, the bagging technique implies the repetitive and random selection of the dataset

k times. Consequently, the trees are all trained using identical information. The end result is shaped from the discrete predictions m_i for every tree within the set by the equation (de Oliveira *et al*. 2021):

$$\hat{m} = \frac{1}{k} \sum_{k}^{i=1} m_i \quad (3.20)$$

3.5.2.3 Boosting Ensemble

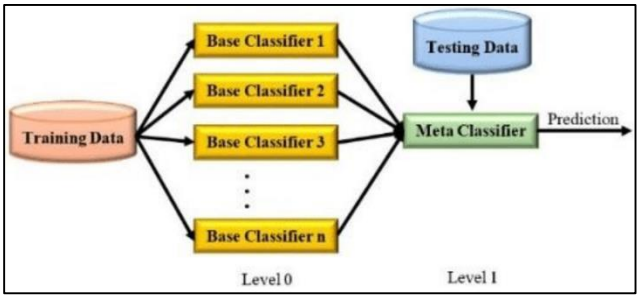
Boosting is a broadly used technique to train poor learners so that they can transform into strong learners. This allows poor learners to classify observations correctly. An incremental tactic is used for those misclassified data points. In order to classify a problem, identically weighted coefficients for all data points are initially used. As the process continues, the weights of the coefficients are reduced for the data points that were classified correctly and augmented for those data points that were misclassified. The tree created during each round decreases errors in the previous round and increases accuracy. This is done by correctly classifying previously misclassified data.

AdaBoost and XGBoost algorithms will be used for classification in this study (Ahmad *et al*., 2020). An AdaBoost classifier begins with a classifier being fit on the initial dataset. Thereafter, additional duplicates are fit on the equivalent dataset, where the incorrectly grouped samples are balanced. The aim is for consequent classifiers to focus more on the problem cases (Hastie *et al*., 2009). XGBoost is an " extreme gradient boosting" ensemble method tailored to be deployed with complicated and large datasets. It is a regularised form of boosting that aids in model generalisation and is also scalable. Moreover, it can also manage distributed and parallel computation, and sparse data effectively, thereby enabling learning to become faster (Chen, 2016).

3.5.2.4 Stacking Ensemble

Stacking usually yields better performance than an individually trained model (Kowsari *et al*., 2018). A stacking model architecture includes at least two sub-models or base models (level-0), and a meta-model (level-1). The meta-model is used to combine predictions of the base models. It is trained on the out-of-fold predictions made during k-fold cross-validation, as illustrated below.

Figure 3.6: Stacking Ensemble Learning (Chanamarn *et al*., 2016, p. 222).



The method is summarised below:

1. Split data into testing and training sets. The training set is then split into k-folds.
2. Fit a sub-model on the (k-1) folds, and make predictions on the kth fold.
3. Iterate the process until all folds have been predicted.
4. Fit the sub-model on the entire training set to compute the performance on the testing set.
5. Iterate steps 2 to 4 for the other sub-models.
6. Utilise predictions on the training set as features in the meta-model.
7. Predict the testing set using the meta-model.

3.6 Chapter Summary

This chapter has presented the methodological steps necessary to realise the second objective of this research study and the associated primary aim. Design research and the associated activities were elaborated on in terms of the study. The proposed framework was discussed with a detailed account of the processes involved at each level. The Liar and ISOT fake news datasets were introduced and their respective configurations highlighted. A variety of performance metrics that will be applied to appraise the machine learning models were also presented. Cross-validation as a resampling method was justified, and the train-test approach and k-fold cross validation-methods were expounded. Finally, the machine learning algorithms selected for Blending Ensemble Model 1 and 2 were presented and elucidated for justification of their applicability. The next chapter

provides the evaluation experiments, the subsequent results, and discussions based on the results when linked to the reviewed literature and the proposed study objectives.

CHAPTER FOUR: PRESENTATION AND DISCUSSION OF RESULTS

4.1 Introduction

This chapter implements the blending ensemble algorithm introduced in Chapter Three. A series of experiments were conducted, and the results combined with their corresponding interpretations are set out below to achieve the third objective of the study. The chapter commences with the strategy employed during the model selection process. This is proceeded by an examination of the experimental results of Blending Ensemble Model 1 and a discussion thereof. The results, analysis and discussion associated with Blending Ensemble Model 2 are then presented. Finally a comparative discussion of the results of the blending ensembles ensures.

4.2 Model Selection

Two blending ensembles were created for the experiments using the blending ensemble algorithm presented in Chapter Three. The first, Blending Ensemble Model 1 (BLD1), was based on traditional machine learning models, whilst the second, Blending Ensemble Model 2 (BLD2), was constructed from ensemble models that encompassed bagging, boosting and stacking machine learning models.

For BLD1, nine machine learning models were observed. The top five performing models on the datasets were selected for inclusion in the ensemble. The selection process incorporated a k-fold cross-validation strategy (k=5) during model evaluation. Models were subsequently ranked according to their cross-validation scores.

Table 4.1 and Table 4.2 display the outcomes of the models for features including n-grams and GloVe word embeddings respectively.

Table 4.1: Cross-validation scores using n-grams.

Model	Cross Validation Score
Linear Support Vector Classifier (SVM)	0.880399
Stochastic Gradient Descent Classifier (SGDC)	0.880291
Logistic Regression Classifier (LR)	0.879425
Linear Discriminant Analysis Classifier (LDA)	0.878884
Ridge Classifier (RC)	0.878191
Passive Aggressive Classifier (PAC)	0.874337
Naive Bayes Classifier (NB)	0.846104
Decision Tree Classifier (DT)	0.709897
K Nearest Neighbour Classifier (KNN)	0.539589

Table 4.2: Cross-validation scores using GloVe.

Model	Cross-Validation Score
Stochastic Gradient Descent Classifier (SGDC)	0.847013
Linear Support Vector Classifier (SVM)	0.839717
Logistic Regression Classifier (LR)	0.839132
K Nearest Neighbour Classifier (KNN)	0.838786
Linear Discriminant Analysis Classifier (LDA)	0.838245
Ridge Classifier (RC)	0.826423
Passive Aggressive Classifier (PAC)	0.823240
Decision Tree Classifier (DT)	0.803776
Naive Bayes Classifier (NB)	0.790959

The following observations can be made from the above tabulated cross-validation scores:

- Four models are common in the top five in both scenarios namely; SGDC, SVM, LR and LDA.

- The top five models in Table 4.1 have scored better than every model ranked in the top five in Table 4.2.

- The difference in performance between the top-ranked and fifth-ranked models in Table 4.1 is 0.002208, which is almost four times less when compared to the equivalently ranked models in Table 4.2 which has a difference in score of 0.008768.

The boxplots in the figures below visually confirm the experimental results generated during the k-fold cross-validation process, which was utilsed in evaluating the models.

Figure 4.1: Boxplot using n-grams.

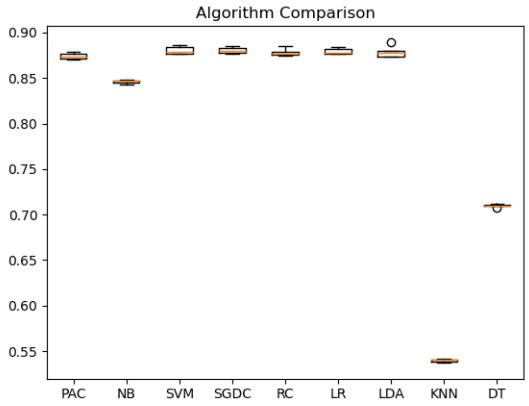
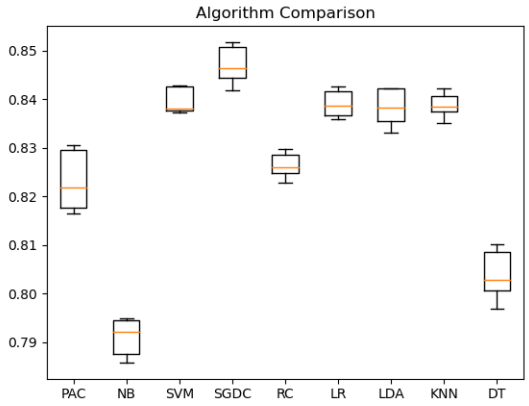


Figure 4.2: Boxplot using GloVe.



Furthermore, it can be observed that the box and whisker plots of the top five performing models in Figure 4.1 are aligned in a straight line, whereas in Figure 4.2, there is a more staggered alignment for the similarly ranked models. This outcome is expected and is in accordance with the observation made earlier relating to the difference between the top and fifth-ranked models according to their cross-validation scores.

4.3 Blending Ensemble Model 1 (BLD1)

BLD1 was constructed using the top five traditional machine learning models that were selected using k-fold cross-validation. The sequence of experimental results is presented separately for GloVe word embeddings and n-grams respectively. For each feature set, the performance metrics namely, receiver operating curve, precision-recall curve and confusion matrix are included for both the Liar and ISOT datasets. Six models and six measurements of performance have been applied in the comparison. These comprise roc auc, auc, f1-score, recall, precision and accuracy. The performance measurements are computed for the fake and real classes.

4.3.1 Performance metrics results obtained from using GloVe word embeddings

Table 4.3 below, summarizes the experiment results on Liar. LDA, SGDC and KNN have peformed the best among the base models. All three models obtained the highest score for two of the six metrics calculated. Although LDA has the best accuracy, SGDC has the best f1-score and KNN the best auc score. Consequently, there is no clear cut winner. However, overall BLD1 has provided the best performance. BLD1 produced the highest scores in four of the six metrics. These include accuracy, auc, roc auc and precision.

Table 4.3: Metrics on Liar using GloVe (BLD1).

	Liar using GloVe					
	roc auc	auc	f1- score	recall	precision	accuracy %
LR	0,600	0,636	0,660	0,740	0,595	58,009
LDA	0,600	0,635	0,662	0,742	0,597	58,274
SGDC	0,597	0,630	0,686	0,850	0,575	57,183
KNN	0,604	0,640	0,673	0,790	0,586	57,759
SVM	0,599	0,635	0,678	0,805	0,586	58,009
BLD1	**0,633**	**0,670**	**0,682**	**0,778**	**0,607**	**60,128**

Table 4.4 below, summarizes the experimental results for the ISOT Dataset using GloVe word embeddings. SGDC and SVM are the top performing sub-models with ISOT having the highest scores in four of six metrics. SGDC may furthermore be considered to be marginally ahead of SVM since it has the better f1-score and accuracy. Overall however, BLD1 is unparalled performance-wise. achieving the five highest scores for roc auc, auc, f1-score, recall, and accuracy.

Table 4.4: Metrics on ISOT using GloVe (BLD1).

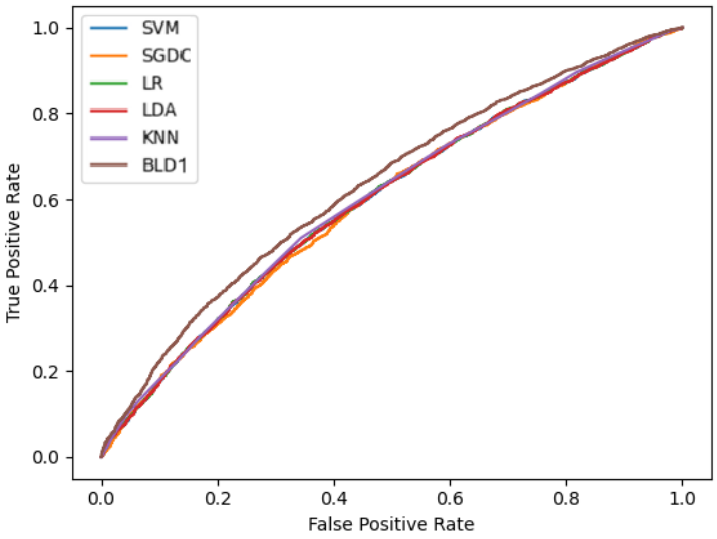
	ISOT using Glove					
	roc auc	auc	f1- score	recall	precision	accuracy %
LR	0,985	0,983	0,942	0,948	0,935	94,418
LDA	0,982	0,980	0,934	0,943	0,926	93,697
SGDC	0,985	0,984	0,944	0,949	0,938	94,592
KNN	0,979	0,976	0,921	0,946	0,899	92,329
SVM	0,986	0,984	0,943	0,949	0,973	94,561
BLD1	**0,987**	**0,986**	**0,947**	**0,953**	**0,941**	**94,891**

4.3.1.1 Receiver Operating Characteristic (ROC) Curve

A classifier's performance can be verified and visualised with the aid of a ROC curve. The curve is produced by plotting the false positive rate and true positive rate on the x and y axes repectively. The region underneath the curve is regarded as a crucial metric in gauging a classifier's performance. It speaks to the "degree or measure of separability" and records performance measurements at a variety of thresholds (Shukla *et al*., 2019). As such, it indicates the skill of a model in differentiating among diverse classes. A high value for ROC will imply that the model's abilty is more effective in distinguishing a true news report from one that is fake.

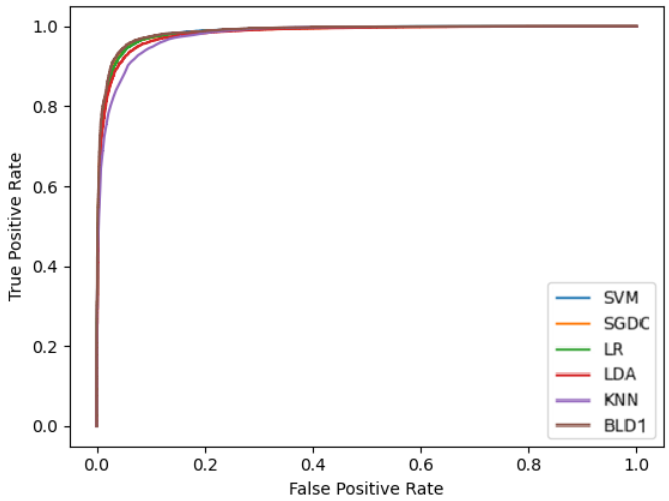
The ROC curves on Liar and ISOT are illustrated below in Figure 4.3 and Figure 4.4 respectively. For both datasets, it can be observed that BLD1 is the superior performer since the region enclosed under the curve is the largest.

Figure 4.3: ROC Curve on Liar using GloVe (BLD1).



In Figure 4.3 we observe that the area under the ROC Curve for SGDC is the smallest, whilst in Figure 4.4, KNN satisfies the same condition. It can thus be concluded that SGDC is the worst performing model on Liar, and KNN is the most unfavourable performing model on ISOT. The inferences can be easily verified by the corresponding scores for roc auc in the metrics tables.

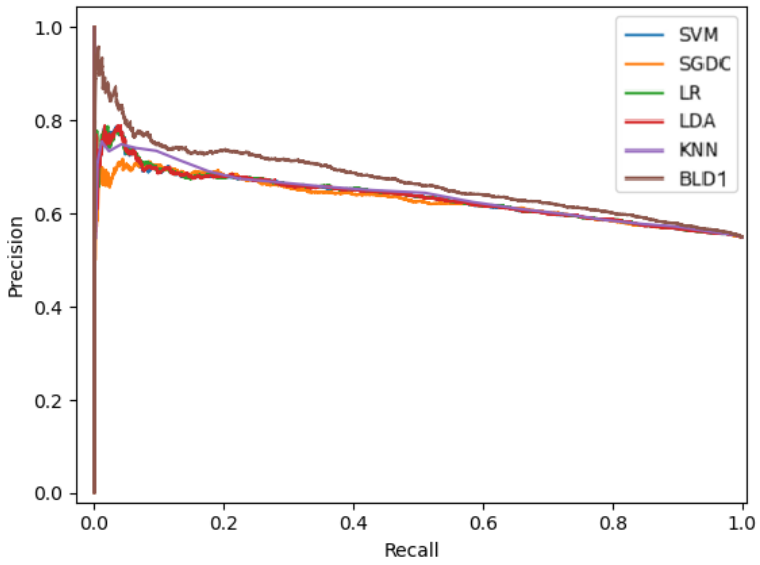
Figure 4.4: ROC Curve on ISOT using Glove (BLD1).



4.3.1.2 Precision-Recall (P-R) Curve

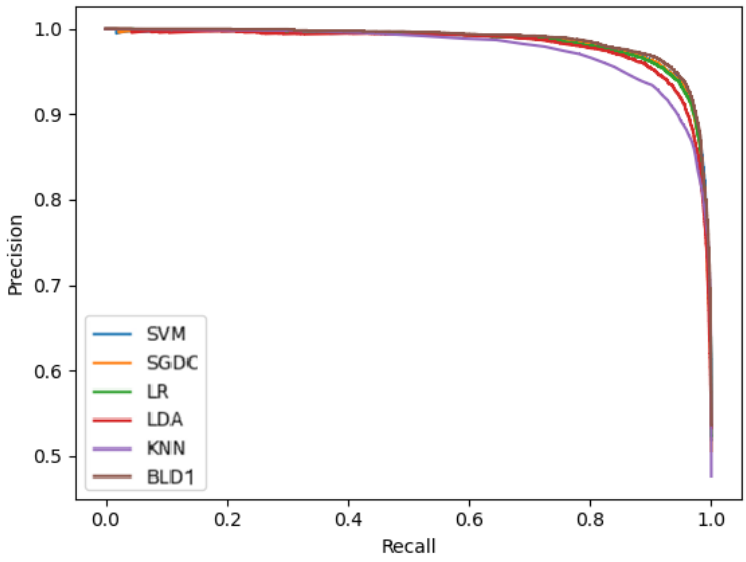
The P-R Curves for the Liar and ISOT datasets are displayed below,in Figure 4.5 and Figure 4.6 respectively. A P-R curve is created by calculating and plotting the recall and and precision values on the x and y axes respectively at different thresholds. It is a summary of the trade-off between the true positive rate and positive prediction label for a prediction model. A good classification model sustains a high precision and high recall all the way through the plot, and will "hug" the corner in the top right hand-side of the graphs below.

Figure 4.5: P-R Curve on Liar using Glove (BLD1).



It is observed in Figure 4.5 that BLD1 is closest to the right upper corner. This indicates that BLD1 has outperformed all the other classifiers and this result is supported by the auc scores in the performance metrics table.

Figure 4.6: P-R Curve on ISOT using Glove (BLD1).

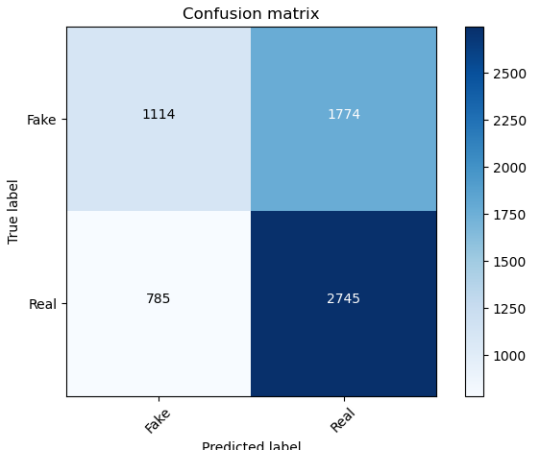


In Figure 4.6 , the plots for all the classifiers are closer to the upper right corner in comparison to Figure 4.5. Hence, it can be concluded that all the classifiers have delivered substantially better performance on ISOT. This conclusion is supported by the scores for auc in the metrics tables. Again, BLD1 is the top performer. It has scored the best for auc on both datasets.

4.3.1.3 Confusion Matrices

These are applied during the performance analysis of a model. A confusion matrix summarises the results of the prediction and consists of the False Negatives, True Positives, True Negatives and False Positives. Figure 4.7 and Figure 4.8 shown below present the confusion matrices of BLD1 on predictions for Liar and ISOT using GloVe respectively.

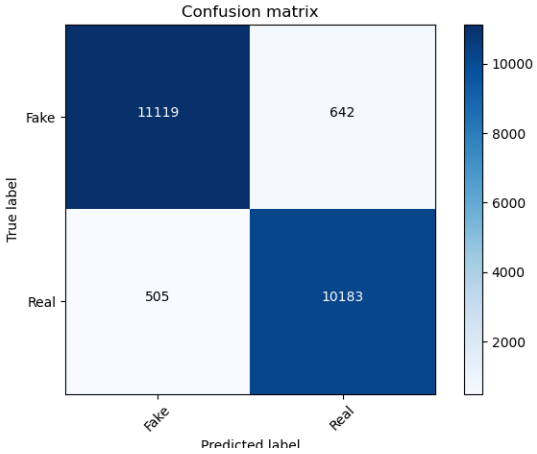
Figure 4.7: Confusion Matrix for Liar using Glove (BLD1).



The following deductions can be drawn from Figure 4.7:

- 1114 fake reports are correctly predicted fake.
- 2745 true reports are correctly predicted true.
- 1774 fake reports are erroneously predicted true.
- 758 true reports are erroneously predicted fake.

Figure 4.8: Confusion Matrix for ISOT using Glove (BLD1).



Similarly, the following inferences can be deduced from Figure 4.8:

- 11119 fake reports are correctly predicted fake.
- 10183 true reports are correctly predicted true.

- 642 fake reports are erroneously predicted true.
- 505 true reports are erroneously predicted fake.

4.3.2 Performance metrics results obtained from using n-grams

Table 4.5 below, summarizes the experimental results on Liar. The superior performing sub-model on Liar is LR, which attained the highest scores of all the sub-models for four of the six metrics. These include roc auc, auc, precision, accuracy. However, although BLD1 achieved the same roc auc and precision scores as LR, BLD1 also delivered the best scores for precision and accuracy making it the overall highest performance model.

Table 4.5: Metrics on Liar using n-grams (BLD1).

	Liar using n-grams					
	roc auc	**auc**	**f1- score**	**recall**	**precision**	**accuracy %**
LR	0,634	0,666	0,673	0,742	0,616	60,346
LDA	0,553	0,596	0,573	0,558	0,589	54,269
SGDC	0,616	0,648	0,682	0,805	0,592	58,725
RC	0,598	0,626	0,685	0,836	0,580	57,682
SVM	0,609	0,642	0,682	0,810	0,589	58,523
BLD1	**0,634**	**0,668**	**0,682**	**0,765**	**0,616**	**60,813**

Table 4.6 below, summarises the experimental results on ISOT. SVM is the top performing sub-model on ISOT scoring the leading scores in five of the six metrics. These include roc auc, auc, f1-score, recall and accuracy. BLD1 achieved the same roc auc score as SVM and a slightly lower auc score but outperformed all other models in the remaining four metrics as such is the top-performing model.

Table 4.6: Metrics on ISOT using n-grams (BLD1).

	ISOT using n-grams					
	roc auc	**auc**	**f1- score**	**recall**	**precision**	**accuracy %**
LR	0,998	0,997	0,979	0,978	0,981	98,036
LDA	0,997	0,997	0,977	0,969	0,986	97,875
SGDC	0,998	0,997	0,982	0,983	0,980	98,254
RC	0,998	0,997	0,980	0,981	0,979	98,071
SVM	0,998	0,998	0,983	0,983	0,983	98,387
BLD1	**0,998**	**0,997**	**0,984**	**0,984**	**0,984**	**98,481**

4.3.2.1 Receiver Operating Characteristic (ROC) Curve

The ROC Curves using n-grams on Liar and ISOT respectively are displayed below in Figure 4.9 and Figure 4.10:

Figure 4.9: ROC Curve on Liar using n-grams (BLD1).

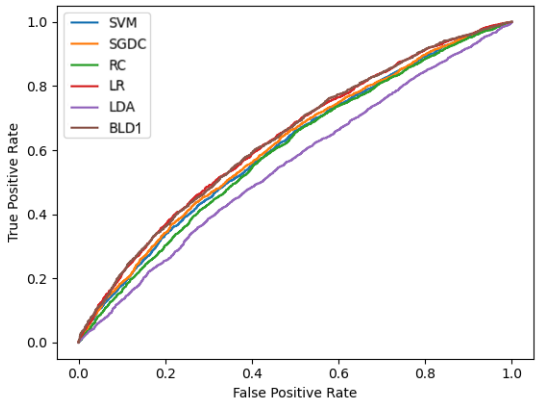
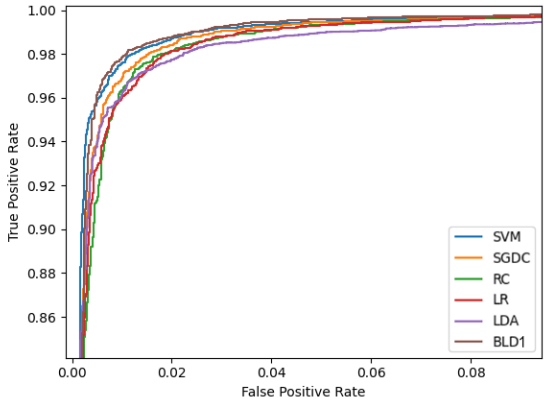


Figure 4.10: ROC Curve on ISOT using n-grams (BLD1).



In both cases, it can be observed that BLD1 is the superior performer because the regions covered under the curves are the largest. Furthermore, it is to be noted that the areas under the ROC Curves for LDA are the smallest. It can consequently be concluded that LDA is the worst performer on both datasets. These deductions can be easily verified by looking up the corresponding scores for roc auc in the metrics tables.

4.3.2.2 Precision-Recall (P-R) Curve

The P-R Curves on Liar and ISOT using n-grams are illustrated below in Figure 4.11 and 4.12 respectively:

Figure 4.11: P-R Curve on Liar using n-grams (BLD1).

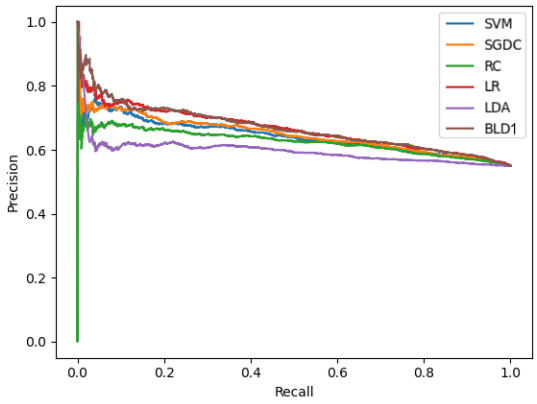
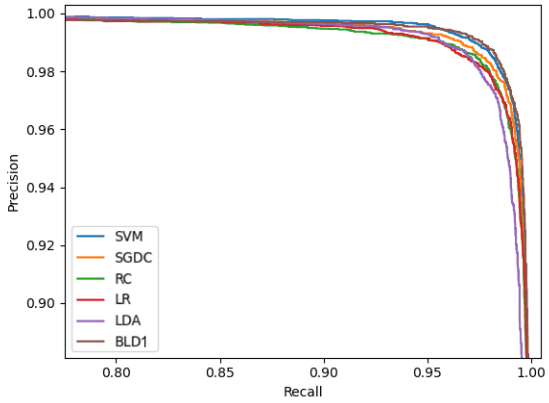


Figure 4.12: P-R Curve on ISOT using n-grams (BLD1).



An effective classifier will "hug" the top right-hand side corner in the graphs. This is apparent for ISOT and is indicative of considerably better performance by every one of the classifiers over Liar. The observation is supported by the scores for auc in the metrics tables. BLD1 appears extremely prominently in the comparison graphs. It has scored the best for auc on Liar and second-best, a mere 0.001 less than SVM, on ISOT.

4.3.2.3 Confusion Matrices

Figure 4.13 and Figure 4.14 shown below present the confusion matrices of BLD1 on predictions for Liar and ISOT using n-grams respectively.

Figure 4.13: Confusion Matrix for Liar using n-grams (BLD1).

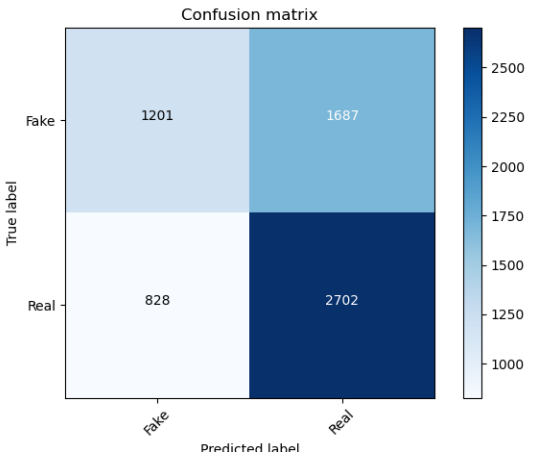


Figure 4.13 yields the following summations:

- 1201 fake reports are correctly predicted fake.
- 2702 true reports are correctly predicted true.
- 1687 fake reports are erroneously predicted true.
- 828 true reports are erroneously predicted fake.

Figure 4.14: Confusion Matrix ISOT using n-grams (BLD1).

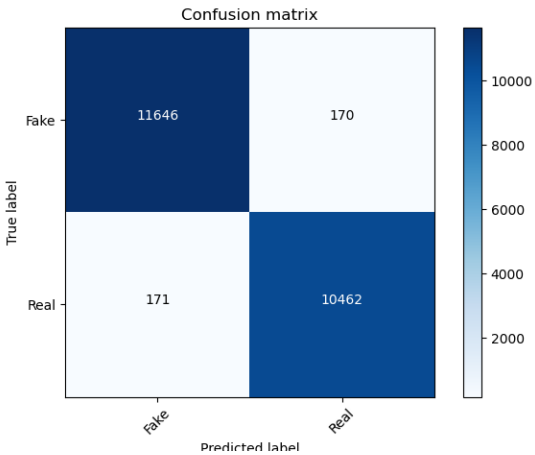


Figure 4.14 summarise the following results:

- 11646 fake reports are correctly predicted fake.
- 10462 true reports are correctly predicted true.
- 170 fake reports are erroneously predicted true.
- 171 true reports are erroneously predicted fake.

4.4 Blending Ensemble Model 2 (BLD2)

BLD2 was constructed from ensemble models, and can be described as 'an ensemble of ensembles'. The models used to build BLD2 include a Stacking Ensemble, Voting Ensemble, Random Forest, AdaBoost and Extreme Gradiant Boosting. The selection includes a mix of bagging, boosting and stacking machine learning models to ensure diversity.

4.4.1 Performance metrics results obtained from using GloVe word embeddings

Table 4.7 below, summarises the experimental result on the Liar dataset using GloVe word embeddings. The top performing sub-model on Liar was STK, which scored the best of the sub-models in four of the six metrics, namely roc auc, auc, precision and accuracy. BLD2 achieved the same auc score as STK but outperformed it on all of the remaining five metrics for the Liar dataset.

Table 4.7: Metrics on Liar using Glove (BLD2).

	Liar using GloVe					
	roc auc	auc	f1- score	recall	precision	accuracy %
Stacking Ensemble (STK)	0,611	0,648	0,673	0,775	0,595	58,616
Voting Ensemble (VS)	0,607	0,644	0,677	0,794	0,590	58,305
Random Forest (RF)	0,592	0,635	0,645	0,713	0,589	56,825
AdaBoost (ADB)	0,574	0,611	0,630	0,684	0,584	55,843
Extreme Gradient Boosting (XGB)	0,576	0,618	0,628	0,677	0,586	55,960
Blending Ensemble (BLD2)	**0,645**	**0,684**	**0,684**	**0,767**	**0,617**	**61,000**

Table 4.8 below, summarises the experimental results for the ISOT dataset. STK was once again the top performing sub-model achieving the highest scores of all sub-models in five of the six metrics. BLD2 achieved the same recall score as STK but outperformed it in the rest of the five metrics. BLD2 is thus the top performing model for the ISOT dataset using GloVe.

Table 4.8. Metrics on ISOT using Glove (BLD2).

	ISOT using GloVe					
	roc auc	auc	f1- score	recall	precision	accuracy %
Stacking Ensemble (STK)	0,988	0,987	0,947	0,954	0,940	94,931
Voting Ensemble (VS)	0,987	0,986	0,944	0,951	0,936	94,592
Random Forest (RF)	0,986	0,985	0,941	0,939	0,942	94,374
AdaBoost (ADB)	0,975	0,972	0,915	0,919	0,910	91,844
Extreme Gradient Boosting (XGB)	0,988	0,987	0,944	0,944	0,944	94,677
Blending Ensemble (BLD2)	**0,989**	**0,988**	**0,950**	**0,954**	**0,947**	**95,265**

4.4.1.1 Receiver Operating Characteristic (ROC) Curve

The ROC Curves on Liar and ISOT are displayed in the figures below:

Figure 4.15: ROC Curve on Liar using Glove (BLD2).

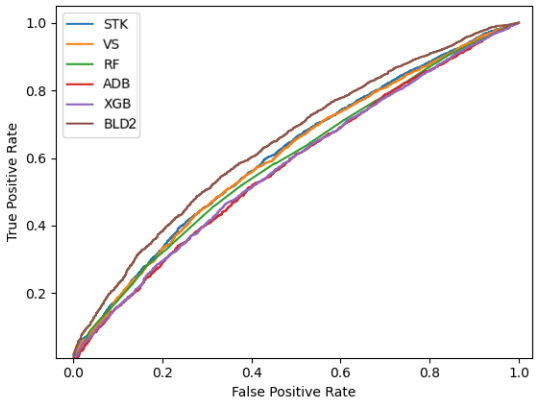
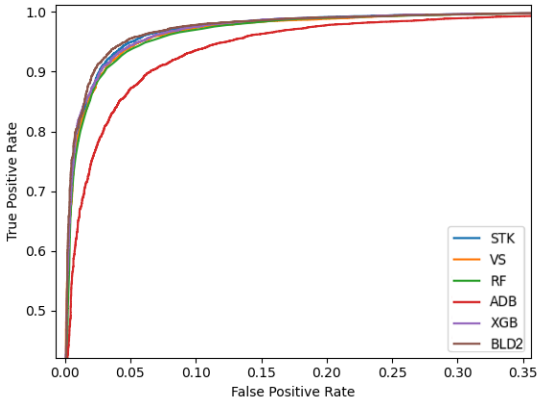


Figure 4.16: ROC Curve on ISOT using Glove (BLD2).



In both scenarious, it can be observed that BLD2 is the most effective model because the areas covered under the curves are the largest. The areas under the ROC Curves for ADB are the smallest. Thus, it can be concluded that ADB is the worst performer on both datasets. These inferences can be confirmed by examining the individual scores for roc auc in the metrics tables.

4.4.1.2 Precision-Recall (P-R) Curve

The P-R Curves on Liar and ISOT for BLD2 using GloVe are displayed in the graphs below:

Figure 4.17: P-R Curve on Liar using Glove (BLD2).

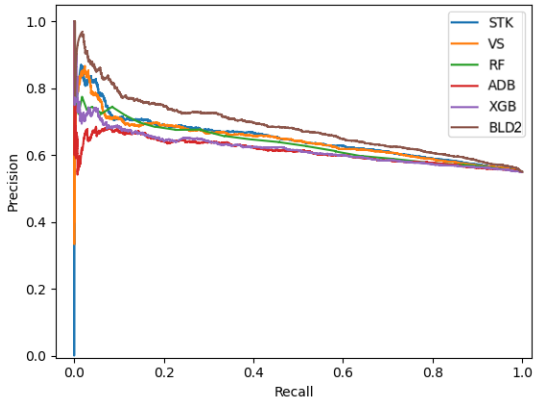
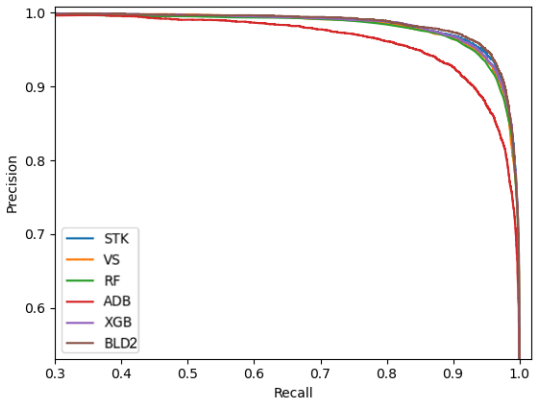


Figure 4.18: P-R Curve on ISOT using Glove (BLD2).



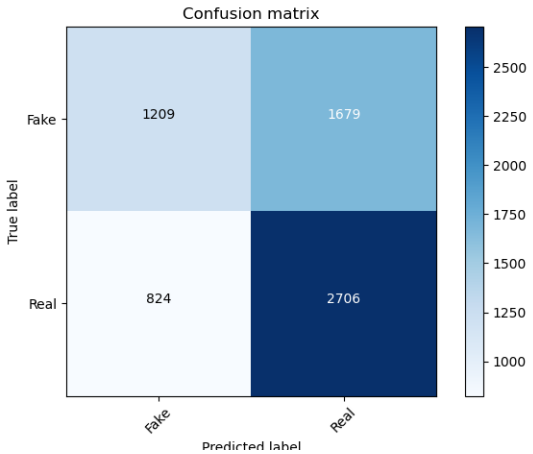
We observe that BLD2 is closest to “hugging” the right upper corner in the plots in Figures 4.17 and 4.18. This is also evident for all the classifiers on the ISOT dataset when compared to the Liar

dataset, and shows significantly better performance by all the classifiers on the ISOT dataset. These observations can be confirmed by the corresponding scores for auc in the metrics tables. BLD2 has the best auc score on both datasets, and is therefore dominant in the comparison plots.

4.4.1.3 Confusion Matrices

Figures 4.19 and 4.20 present the confusion matrices of BLD2 on predictions for Liar and ISOT using GloVe.

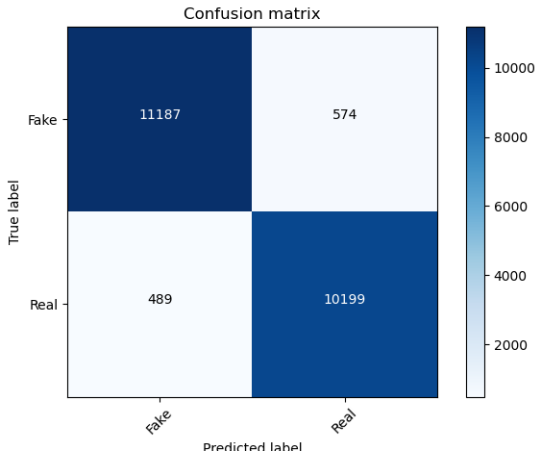
Figure 4.19: Confusion Matrix for Liar using Glove (BLD2).



The following statistics can be gleaned from Figure 4.19:

- 1209 fake reports are correctly predicted fake.
- 2706 true reports are correctly predicted true.
- 1679 fake reports are erroneously predicted true.
- 824 true reports are erroneously predicted fake.

Figure 4.20: Confusion Matrix for ISOT using Glove (BLD2).



Correspondingly, the following conclusions can be drawn from Figure 4.20

- 11187 fake reports are correctly predicted fake.
- 10199 true reports are correctly predicted true.
- 574 fake reports are erroneously predicted true.
- 489 true reports are erroneously predicted fake.

4.4.2 Performance metrics results obtained from using n-grams

Table 4.9 below, summarizes the experiment results on Liar. The top performing sub-model on Liar is STK having scored the best in four of the six metrics. These are for auc, roc auc, accuracy and precision. Overall BLD2 has performaned the best. The four highest scores achieved of the six are auc, roc auc, accuracy and precision.

Table 4.9: Metrics on Liar using n-grams (BLD2).

	Liar using n-grams					
	roc auc	f1- score	auc	precision	recall	accuracy %
STK	0,636	0,679	0,667	0,613	0,761	60,471
VS	0,598	0,620	0,639	0,599	0,643	56,638
RF	0,616	0,660	0,648	0,606	0,724	58,975
ADB	0,589	0,696	0,625	0,582	0,867	58,414
XGB	0,606	0,686	0,643	0,589	0,823	58,648
BLD2	**0,657**	**0,686**	**0,692**	**0,621**	**0,766**	**61,483**

Table 4.10 below summarises the experiment results using n-grams on the ISOT dataset. STK and VS were the top performing sub-models on ISOT displaying the highest scores in four of the six metrics. Furthermore, STK may be considered to be marginally ahead of VS since it has the better accuracy score and is just 0.001 behind in terms of the auc score. Overall BLD2 is the superior performing model having yielded the most impressive score in every performance metric category.

Table 4.10: Metrics on ISOT using n-grams (BLD2).

	ISOT using ngram					
	roc auc	auc	f1- score	recall	precision	accuracy %
STK	0,998	0,997	0,983	0,982	0,983	98,383
VS	0,998	0,998	0,983	0,981	0,985	98,379
RF	0,996	0,996	0,972	0,980	0,964	97,283
ADB	0,981	0,983	0,932	0,905	0,960	93,715
XGB	0,988	0,988	0,938	0,908	0,970	94,280
BLD2	**0,999**	**0,998**	**0,986**	**0,987**	**0,985**	**98,641**

4.4.2.1 Receiver Operating Characteristic (ROC) Curve

The ROC curves on Liar and ISOT datasets using n-grams are illustrated in the figures below:

Figure 4.21: ROC Curve on Liar using n-grams (BLD2).

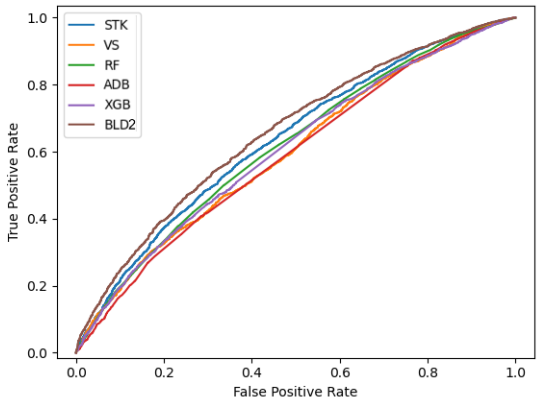
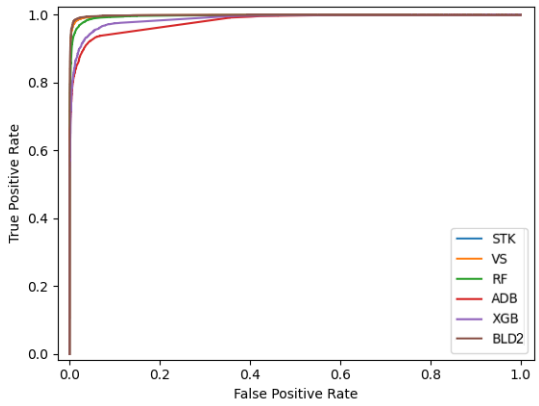


Figure 4.22: ROC Curve on ISOT using n-grams (BLD2).



For both cases, it is clear that BLD2 is the best model because the regions covered under the curves are the greatest. In addition, the areas under the ROC Curves for ADB are the smallest. Hence, it can be concluded that ADB is the worst performer on both datasets. These inferences can be corroborated easily by checking the corresponding roc auc score in the metrics tables.

4.4.2.2 Precision-Recall (P-R) Curve

The P-R Curves on Liar and ISOT using n-grams are displayed in the graphs below:

Figure 4.23: P-R Curve on Liar using n-grams (BLD2).

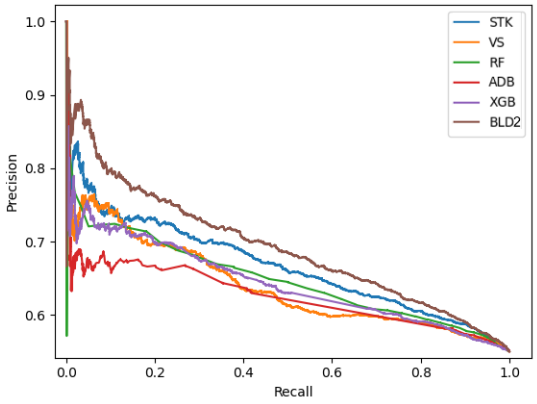
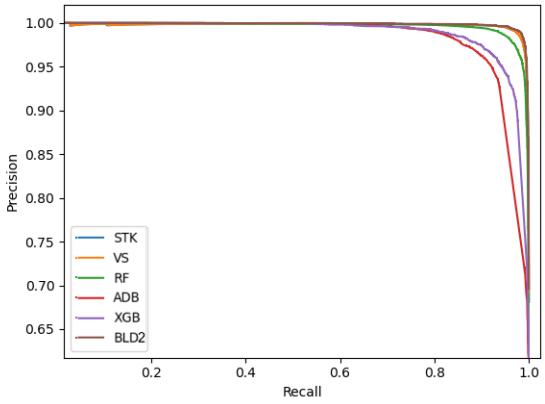


Figure 4.24: P-R Curve on ISOT using n-grams (BLD2).

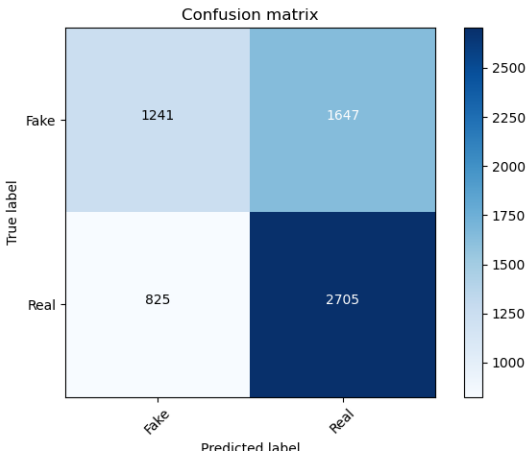


We notice that BLD2 is once again closest to "hugging" the upper right-hand side corner in both the graphs. This is applicable for every classifier on ISOT over Liar, and indicates significantly superior performance by all classifiers on the ISOT dataset. These observations can be verified by the scores for auc in the metrics tables. BLD2 has the highest auc score on both datasets and is therefore leading in the comparison plots.

4.4.2.3 Confusion Matrices

Figures 4.25 and 4.26 illustrate the confusion matrices of BLD2 on predictions using n-grams for Liar and ISOT respectively.

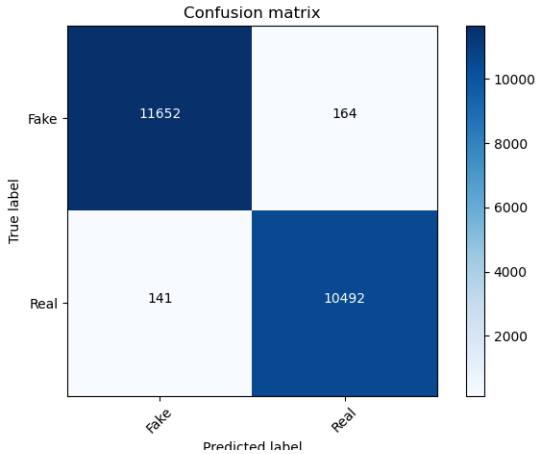
Figure 4.25: Confusion Matrix for Liar using n-grams (BLD2).



The following interpretations can be made from Figure 4.25:

- 1241 fake reports are correctly predicted fake.
- 2705 true reports are correctly predicted true.
- 1647 fake reports are erroneously predicted true.
- 825 true reports are erroneously predicted fake.

Figure 4.26: Confusion Matrix for ISOT using n-grams (BLD2).



Correspondingly, the following information can be extrapolated from Figure 4.26:

- 11652 fake reports are correctly predicted fake.
- 10492 true reports are correctly predicted true.
- 164 fake reports are erroneously predicted true.
- 141 true reports are erroneously predicted fake.

4.5 Comparison of the Blending Ensembles

The performance metrics on Liar and ISOT for BLD1 and BLD2 are compared below:

4.5.1 Metrics on Liar

Tables 4.11 and 4.12 illustrate the experiment results on Liar using GloVe and n-grams:

Table 4.11: Metrics on Liar Dataset using GloVe.

	Liar using GloVe					
	roc auc	auc	f1- score	recall	precision	accuracy %
BLD1	0,633	0,670	0,682	0,778	0,607	60,128
BLD2	**0,645**	**0,684**	**0,684**	**0,767**	**0,617**	**61,000**

Table 4.12: Metrics on Liar using n-grams.

	Liar using n-grams					
	roc auc	auc	f1- score	recall	precision	accuracy %
BLD1	0,634	0,668	0,682	0,765	0,616	60,813
BLD2	**0,657**	**0,692**	**0,686**	**0,766**	**0,621**	**61,483**

BLD1 only achieved one recall score that was higher than BLD2 across both feature sets. Overall, BLD2 is the superior model on the Liar dataset and was unbeatable when using both GloVe and n-grams, as confirmed by all five of six scores in Table 4.11 and six of six scores in Table 4.12.

4.5.2 Metrics on ISOT

Table 4.13 and Table 4.14 below below illustrate the experiment results on ISOT using GloVe and n-grams respectively.

Table 4.13: Metrics on ISOT using GloVe.

	ISOT using GloVe					
	roc auc	auc	f1- score	recall	precision	accuracy %
BLD1	0,987	0,986	0,947	0,953	0,941	94,891
BLD2	**0,989**	**0,988**	**0,950**	**0,954**	**0,947**	**95,265**

Table 4.14: Metrics ISOT using n-grams.

	ISOT using n-grams					
	roc auc	auc	f1- score	recall	precision	accuracy %
BLD1	**0,998**	**0,997**	**0,984**	**0,984**	**0,984**	**98,481**
BLD2	**0,999**	**0,998**	**0,986**	**0,987**	**0,985**	**98,641**

It is evident that BLD2 achieved the top performance scores for every metric applied using both GloVe and n-grams. It can thus be concluded that BLD2 is by far the superior model on the ISOT dataset and was unconquerable.

4.6 Discussion of Results

The sections below summarise the results for each model and posit likely reasons for these results in terms of the literature reviewed in Chapter Two and the methodological framework presented in Chapter Three. These are followed by a summary of the comparison of the results produced by experimentations on the two models.

4.6.1 Summary of results for BLD1

BLD1 has outclassed all the base models used in the ensembles for both GloVe and n-grams as features. The result was reasonably anticipated as the model was based on various successful studies involving ensemble machine learning, as cited in the review of literature included in Chapter Two. However, there seems to be big gap between the corresponding performance metrics on the Liar dataset and ISOT dataset. This is largely attributable to the Liar dataset comprising of mostly short statements or sentences, and the relatively smaller size of the dataset when compared to the ISOT dataset, which is made up of full texts. Nevertheless, the performance of every model on the Liar dataset is comparable to the best in many previous studies, including benchmark studies, and is also an improvement. These include studies by Agarwal *et al*., (2019), Khan *et al*., (2019), Shukla *et al*., (2019), and Wang (2017).

Finally, n-grams as features have delivered better performance compared to GloVe by a fair margin. This result is compatible with those that have been realised in similar projects. Vijayaraghavan *et al*., (2020) applied Word2Vect embeddings and experimentally demonstrated that it performed the worst when compared to models using TF-IDF. Thota *et al*., (2018) produced superior performance with n-grams when compared to word embeddings. Analogous findings were also reported by Smitha and Bharath (2020).

4.6.2 Summary of results for BLD2

BLD2 has outshone all the base models used in the ensembles for both GloVe and n-grams features. The result was also highly predictable as it has been confirmed in several studies

involving ensemble machine learning as stated in Chapter Two. N-grams as features have delivered superior performance in comparison to GloVe once again, and there is also quite a gap between the corresponding performance metrics on the Liar dataset and ISOT dataset. The reasons and explanation associated to these results are analogous to what has been expressed in the discussion under section 4.6.1.

4.6.3 Summary of comparison

BLD2 was the top performer on the Liar and ISOT datasets across both GloVe and n-grams as features. The results have been validated by applying a combination of an assortment of metrics used to quantify performance. The best performance by BLD2 was recorded using n-grams on the ISOT dataset. As such, this is the model that will be saved for future use. The results of the blending ensemble compare very favourably with results presented in previous studies, and this includes the benchmark studies. More specifically, it is an overall performance enhancement in comparison to several of the ensembles presented by Ahmad *et al*., (2020) on the ISOT dataset, and this also includes the two benchmark models, Wang-Bi-LSTM and Wang–CNN.

4.7 Chapter Summary

The blending ensemble algorithm presented in Chapter Three was implemented in creating Blending Ensemble Model 1 and Model 2. This chapter has presented the evaluation experiments, results and discussions for both models. The experiments were conducted using two benchmark fake news datasets, and included GloVe and n-grams as features. Six well known statistical metrics were applied during the evaluation of the models. ROC Curves, P-R Curves and Confusion Matrices were also included to add a visual dimension to the interpretation and appraisal of the models. A total of twenty-four machine learning models were created and tested. Finally, the performance of each model was compared. The experimental results of the blending ensemble models are very encouraging, and represent an improvement on many previous studies including benchmark studies. The next chapter will provide the conclusion to this dissertation.

CHAPTER FIVE: SUMMARY AND CONCLUSIONS

5.1 Introduction

This chapter is the concluding part of this dissertation and includes an overview of the study conducted, contribution to knowledge and possible research directions for further improvements based upon the proposed ensemble approach.

5.2 Summary

In this study, a blending ensemble algorithm for fake news detection was investigated, created, implemented, and then tested with the aim of providing a feasible solution that will improve the existing methods for the detection of fake news. The overarching goal of the study was to detect fake news by applying an ensemble approach to combat the proliferation and propagation of fake news. Based on the results of the study, the researcher believes he has accomplished his goal. As such, the research objectives have all been satisfied so as to accomplish the aim of this study. They are restated below with their accompanying conclusions for the sake of clarity:

1. To comprehensively research relevant publications based on the detection of fake news to identify methods and approaches in devising a probable solution.

The outcome of the first objective was presented in Chapter Two with an in-depth review of the most current literature available to support and motivate for the necessity of this study. Chapter Two presented a survey of fake news, the combatting strategies, detection algorithms and meta-analysis of the efficacy of existing studies. Current trends in tackling the challenging menace of fake news were also highlighted. The emphasis was strongly focused on the different methods and strategies proposed in the literature by recognising their associated strengths and drawbacks. The detection of fake news is a complex problem, and despite the development and implementation of numerous potential solutions, the problem continues to gain traction at an increasing rate. The dissemination of fake news in various formats presents significant threats to society at large. This served as motivation to develop an advanced ensemble approach to improve the detection of fake news.

2. To develop an ensemble approach to detect fake news by combining machine learning with natural language processing.

The methodological steps that were necessary to realise the second objective of this research study were presented in Chapter Three. A design research strategy that included both the power of natural language processing and machine learning was adopted. A blending ensemble algorithm was developed to bolster the automated detection of fake news. Two benchmark datasets were identified, and the feature extraction process was clarified. An assortment of machine learning models was presented for inclusion in Blending Ensemble Model 1 and Model 2.

3. To experimentally evaluate the developed model against other machine learning algorithms using well-known statistical evaluation metrics.

In Chapter Four, the third objective was intrinsically met by experimentally validating the performance of the proposed blending ensemble algorithm against other machine learning algorithms using well-known statistical evaluation metrics. This chapter presented an analysis of the testing results achieved for Blending Ensemble 1 and Blending Ensemble 2 and discussed in terms of a comparison with the results obtained by five existing models that were tested on the same two datasets. The experimentation was conducted on a pair of benchmark fake news datasets and included GloVe and n-grams as features. Six well-known statistical metrics were applied during the evaluation process. Furthermore, the inclusion of ROC Curves, P-R Curves and Confusion Matrices added a visual layer to the interpretation and evaluation of the models. A total of twenty-four machine learning models were built and deployed during experimentation. The experimental results of the blending ensembles are very promising, particularly for Blending Ensemble Model 2, and represent an improvement over many previous studies including benchmark studies. More importantly, the research reported in this study has confirmed the efficacy of the newly proposed blending ensemble algorithm to improve the detection of fake news.

5.3 Research Contributions

It is the researcher's intention that the proposed dissertation will aid in filling the gap in the area of study by proposing a new ensemble approach that incorporates a blended ensemble algorithm, with the aim of more accurately detecting fake news. The primary research question of what ensemble approach could be developed to ameliorate the detection of fake news is targeted towards the development of a new blending ensemble algorithm in this study. The main contributions of the new blending ensemble algorithm developed in this study are envisaged to:

(a) Demonstrate the efficacy of the newly proposed blending algorithm to detect fake news.

(b) Create a reliable quantitative method as a means to evaluate the proposed blending ensemble algorithm quantitatively by consuming fake news articles acquired from two publicly available benchmark datasets, and applying six well-known statistical evaluation metrics in terms of precision, recall, accuracy, roc, auc, and f1-score.

(c) Compare performance evaluation against other machine learning models to demonstrate the efficiency and superiority of the proposed blending ensemble algorithm.

(d) Contribute to the machine learning field by presenting the blending ensemble model as an effective solution for detecting fake news thereby improving upon existing approaches.

5.4 Future Work

Despite significant research progress having been achieved in the detection of fake news, there will always be room for innovation and improvement. As such, the researcher has outlined potential extensions and applications by highlighting a few of the many exciting future works related to the blending ensemble algorithm developed in this study. In the view of the researcher, the following studies are worth pursuing:

1. Experimentation on other datasets from different social contexts to investigate the impact of varying the configuration of the ensemble.

2. The online ecosystem and social media, in particular, are shaping our lifestyles. It is imperative then to take into consideration current trends in social media so that these can be integrated with the blending ensemble model to enhance detection of fake news.

3. Investigate the prospect of creating feature ensembles by adapting the blending ensemble algorithm.

4. The proposed blending algorithm can also be extended and applied to other practical application domains such as the detection of fake images and audio.

5.5 Conclusion

The battle against fake news is a dynamic area of research with many gaps yet to be plugged. This study has acknowledged that despite numerous research efforts, fake news is still a burgeoning problem that continues to adversely impact our lives on both a local and global scale. This dissertation investigated the problem of fake news and tackled the task of attempting to improve upon existing automated methods of fake news recognition. More specifically, as a result of the gap identified through a comprehensive review of current studies and literature on the topic, the researcher chose to develop an ensemble stategy for detecting fake news so as to combat its proliferation and propagation. The work reported in this study has investigated this novel blending ensemble approach by combining the processing of natural language and machine learning towards more efficient fake news identification.

The proposed blending ensemble was tested on two benchmark fake news datasets and measured against existing models using a variety of metrics to gauge efficacy and validate performance. The blending ensemble algorithm was demonstrated to outclass most other existing benchmark algorithms. Therefore, it can be concluded that the blending ensemble model is indeed a feasible solution for boosting fake news detection. Finally, the research emanating from this study has been acknowledged, and already published in an international journal titled, *Scientific Programming* in the form of an article entitled; “Detection of Online Fake News Using Blended Ensemble Learning” (Hansrajh *et al*., 2021).

REFERENCES

Absa (2021). Protect yourself from phishing scams. [online] Available at: <https://www.absa.co.za/self-service/safety-security/phishing-scams/> [Accessed 23 June 2021].

Agarwal, A. and Dixit, A. (2020). Fake News Detection: An Ensemble Learning Approach. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE.

Agarwal, V., Sultana, H. P., Malhotra, S. & Sarkar, A. (2019). "Analysis of Classifiers for Fake News Detection." Procedia Computer Science 165: 377-383.

Ahmad, I., Yousaf, M., Yousaf, S. & Ahmad, M. O. (2020). "Fake News Detection Using Machine Learning Ensemble Methods." Complexity 2020: 8885861.

Ahmed, H., Traore, I. & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. International conference on intelligent, secure, and dependable systems in distributed and cloud environments, Springer.

Ahmed, H., Traore, I. & Saad, S. (2018). "Detecting opinion spams and fake news using text classification." Security and Privacy 1(1): e9.

Ahmed, S., Hinkelmann, K. & Corradini, F. (2019). Combining machine learning with knowledge engineering to detect fake news in social networks-a survey. Proceedings of the AAAI 2019 Spring Symposium.

Ajao, O., Bhowmik, D. & Zargari, S. (2018). Fake news identification on twitter with hybrid cnn and rnn models. Proceedings of the 9th international conference on social media and society.

Aker, A., Derczynski, L. & Bontcheva, K. (2017). "Simple open stance classification for rumour analysis." arXiv preprint arXiv:1708.05286.

Al-Ash, H. S., Putri, M. F., Mursanto, P. & Bustamam, A. (2019). Ensemble Learning Approach on Indonesian Fake News Classification. 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), IEEE.

Alguliyev, R. M., R. Aliguliyev, M. & Sukhostat, L. V. (2020). "Weighted consensus clustering and its application to Big data." Expert Systems with Applications 150: 113294.

Allcott, H .& Gentzkow, M. (2017). "Social Media and Fake News in the 2016 Election." Journal of Economic Perspectives 31(2): 211-236.

Aphiwongsophon, S. & Chongstitvatana, P. (2018). Detecting fake news with machine learning method. 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), IEEE.

Bajaj, A. (2021). Performance Metrics in Machine Learning [Complete Guide] - neptune.ai. [online] Available at: <https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide> [Accessed 14 July 2021].

BBC News (2021). Coronavirus: The viral rumours that were completely wrong. [online] Available at: <https://www.bbc.com/news/blogs-trending-53640964> [Accessed 23 June 2021].

Bickham, A., Howard, C. & Simmons, S. L. (2018). "Overview of Fake News: For Public Organizations." [online] Available at: < https://stars.library.ucf.edu/ > [Accessed 20 June 2021].

Biyani, P., Tsioutsiouliklis, K. & Blackmer, J. (2016). "8 amazing secrets for getting more clicks": detecting clickbaits in news streams using article informality. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona, AAAI Press: 94–100.

Bodnar, T., Tucker, C., Hopkinson, K. & Bilén, S. G. (2014). Increasing the veracity of event detection on social media networks through user trust modeling. 2014 IEEE International Conference on Big Data (Big Data), IEEE.

Bolón-Canedo, V. & Alonso-Betanzos, A. (2019). "Ensembles for feature selection: A review and future trends." Information Fusion 52: 1-12.

Bondielli, A. & Marcelloni, F. (2019). "A survey on fake news and rumour detection techniques." Information Sciences 497: 38-55.

Brandtzaeg, P. B. & Følstad, A. (2017). "Trust and distrust in online fact-checking services." Communications of the ACM 60(9): 65-71.

Breiman, L., Friedman, J. & Olshen, R. (2017). "Classification and regression trees Routledge." [online] Available at: < https://doi.org/10.1201/9781315139470> [Accessed 12 June 2021].

Briscoe, E. J., Appling, D. S. & Hayes, H. (2014). Cues to deception in social media communications. 2014 47th Hawaii international conference on system sciences, IEEE.

Brownlee, J. (2018). A Gentle Introduction to k-fold Cross-Validation. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/k-fold-cross-validation/> [Accessed 12 July 2021].

Brownlee, J. (2020). Train-Test Split for Evaluating Machine Learning Algorithms. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/> [Accessed 12 July 2021].

Brummette, J., DiStaso, M., Vafeiadis, M. & Messner, M. (2018). "Read all about it: The politicization of “fake news” on Twitter." Journalism & Mass Communication Quarterly 95(2): 497-517.

Burkhardt, J. M. (2017). "History of fake news." Library Technology Reports 53(8): 5-9.

Busari, S. & Adebayo, B. (2020). "Nigeria records chloroquine poisoning after Trump endorses it for coronavirus treatment." CNN. [online] Available at: <https://www. cnn. com/2020/03/23/africa/chloroquine-trump-nigeria-intl/index. html> [Accessed 2 April 2020].

Campan, A., Cuzzocrea, A. & Truta, T. (2017). "Fighting fake news spread in online social networks: Actual trends and future research directions." 2017 IEEE International Conference on Big Data (Big Data): 4453-4457.

Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E. & Freire, J. (2019). A topic-agnostic approach for identifying fake news pages. Companion proceedings of the 2019 World Wide Web conference.

Chakraborty, A., Paranjape, B., Kakarla, S. & . Ganguly, N (2016). Stop clickbait: Detecting and preventing clickbaits in online news media. 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), IEEE.

Chanamarn, N., Tamee, K. & Sittidech, P. (2016). Stacking Technique for Academic Achievement Prediction. 2016 International Workshop on Smart Info-Media Systems in Asia.

Chauhan, N., (2020). Model Evaluation Metrics in Machine Learning - KDnuggets. [online] KDnuggets. Available at: <https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html> [Accessed 14 July 2021].

Chen, Y. C., Liu, Z. Y. & Kao, H. Y. (2017). Ikm at semeval-2017 task 8: Convolutional neural networks for stance detection and rumor verification. Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017).

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078.

Choudhary, A. & Arora, A. (2021). "Linguistic feature based learning model for fake news detection and classification." Expert Systems with Applications 169: 114171.

Chua, A. Y. & Banerjee, S. (2016). Linguistic predictors of rumor veracity on the internet. Proceedings of the International MultiConference of Engineers and Computer Scientists.

Chung, M. & Kim, N. (2021). "When I learn the news is false: How fact-checking information stems the spread of fake news via third-person perception." Human Communication Research 47(1): 1-24.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F. & Flammini, A. (2015). "Computational fact checking from knowledge networks." PloS one 10(6): e0128193.

Cohen, E. (2014). Ebola airborne: A nightmare that could happen - CNN. [online] CNN. Available at: https://edition.cnn.com/2014/09/12/health/ebola-airborne/ [Accessed 23 June. 2021].

Colgan, W. & Kakkar, K. (2019). "Multi-Perspective Ensemble for Hyper-Partisan News Detection."

Collins, B., Hoang, D. T., Nguyen, N. T. & Hwang, D. (2020). "Trends in combating fake news on social media–a survey." Journal of Information and Telecommunication: 1-20.

Conroy, N. K., Rubin, V. L. & Chen,Y. (2015). "Automatic deception detection: Methods for finding fake news." Proceedings of the Association for Information Science and Technology 52(1): 1-4.

Crawford, M., Khoshgoftaar, T . M., Prusa, J. D., Richter, A. N. & Al Najada, H. (2015). "Survey of review spam detection using machine learning techniques." Journal of Big Data 2(1): 1-24.

Data Science Central (2019). Cross Validation in One Picture. [online] Available at: <https://www.datasciencecentral.com/profiles/blogs/cross-validation-in-one-picture> [Accessed 26 July 2021].

Davidson, T., Warmsley, D., Macy, M. & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media.

de Beer, D. & Matthee, M. (2020). Approaches to Identify Fake News: A Systematic Literature Review. International Conference on Integrated Science, Springer.

de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V. & Mattos, D. M. (2021). "Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges." Information 12(1): 38.

Della Vedova, M. L., Tacchini, E., Moret, S., Ballarin,G., DiPierro, M. & de Alfaro, L. (2018). Automatic online fake news detection combining content and social signals. 2018 22nd conference of open innovations association (FRUCT), IEEE.

Ellis, T. J. & Levy, Y. (2010). A guide for novice researchers: Design and development research methods. Proceedings of Informing Science & IT Education Conference (InSITE), Citeseer.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G. & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. Proceedings of the International AAAI Conference on Web and Social Media.

Enayet, O. & El-Beltagy, S. R. (2017). NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter. Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017).

FactCheck (2021). Our Mission - FactCheck.org. [online] Available at: <https://www.factcheck.org/about/our-mission/> [Accessed 28 May 2021].

Fallman, D. (2007). "Why research-oriented design isn't design-oriented research: On the tensions between design and research in an implicit design discipline." Knowledge, Technology & Policy 20(3): 193-200.

Faste, T. & Faste, H. (2012). Demystifying "design research": Design is not research, research is design. IDSA education symposium.

Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. (2014). "Do we need hundreds of classifiers to solve real world classification problems?" The journal of machine learning research 15(1): 3133-3181.

Ferreira, W. & Vlachos, A. (2016). Emergent: a novel data-set for stance classification. Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies.

Figueira, Á. & Oliveira, L. (2017). "The current state of fake news: challenges and opportunities." Procedia Computer Science 121: 817-825.

Flores-Saviaga, C., Keegan, B. & Savage, S. (2018). Mobilizing the trump train: Understanding collective action in a political trolling community. Proceedings of the International AAAI Conference on Web and Social Media.

Full Fact (2021). The media must stop using misleading headlines - Full Fact. [online] Available at: <https://fullfact.org/news/edlines-headlines-that-contradict-the-article/> [Accessed 23 June 2021].

Garg, H., Goyal, A. & Joshi, M. A. (2020). "Techniques of Fake News Detection." Journal of Advanced Research in Instrumentation and Control Engineering 7(2): 8-11.

Gartner (2017). Top Predictions for IT Organizations and Users in 2018 and Beyond. [online] Available at: < www.gartner.com/en/newsroom/press-releases/2017-10-03-gartner-reveals-top-predictions-for-it-organizations-and-users-in-2018-and-beyond> [Accessed 33 June 2021].

Ghosal, D., Bhatnagar, S., Akhtar, M. S., Ekbal, A. & Bhattacharyya, P. (2017). IITP at SemEval-2017 task 5: an ensemble of deep learning and feature based models for financial sentiment analysis. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).

Giasemidis, G., Singleton, C., Agrafiotis, I., Nurse, J. R., Pilgrim, A., Willis, C. & Greetham, D. V. (2016). Determining the veracity of rumours on Twitter. International Conference on Social Informatics, Springer.

Gilda, S. (2017). Evaluating machine learning algorithms for fake news detection. In Research and Development (SCOReD), 15th Student Conference, IEEE.

Granik, M. & Mesyura V. (2017). Fake news detection using naive Bayes classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), IEEE.

Granik, M., Mesyura, V. & Yarovyi, A. (2018). Determining fake statements made by public figures by means of artificial intelligence. 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), IEEE.

Granik, M. O. & Mesyura, V. I. (2018). "FAKE STATEMENTS DETECTION WITH ENSEMBLE OF MACHINE LEARNING ALGORITHMS."

Gutierrez-Espinoza, L., F., Namin, A. S., Jones, K. S. & Sears, D. R. (2020). "Fake Reviews Detection through Ensemble Learning." arXiv preprint arXiv:2006.07912.

Hamidian, S. & Diab, M. T. (2019). "Rumor detection and classification for twitter data." arXiv preprint arXiv:1912.08926.

Hansrajh, A., Adeliyi, T. T. & Wing, J. (2021). "Detection of Online Fake News Using Blending Ensemble Learning." Scientific Programming 2021: 3434458.

Hardalov, M., Koychev, I. & Nakov, P. (2016). In search of credible news. International conference on Artificial intelligence: methodology, systems, and applications, Springer.

Harlow, S. (2015). "Story-chatterers stirring up hate: Racist discourse in reader comments on US newspaper websites." Howard Journal of Communications 26(1): 21-42.

Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J. & Yu, C. (2015). The quest to automate fact-checking. Proceedings of the 2015 Computation+ Journalism Symposium.

Hassan, N., Arslan, F., Li, C. & Tremayne, M. (2017). Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Hassan, N., Yousuf, M., Mahfuzul Haque, M., Suarez Rivas, J. A. & Khadimul Islam, M. (2019). Examining the roles of automation, crowds and professionals towards sustainable fact-checking. Companion Proceedings of The 2019 World Wide Web Conference.

Hastie, T., Rosset, S., Zhu, J. & . Zou, H (2009). "Multi-class adaboost." Statistics and its Interface 2(3): 349-360.

Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004). "Design science in information systems research." MIS quarterly: 75-105.

Hine, G., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G. & Blackburn, J. (2017). Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. Proceedings of the International AAAI Conference on Web and Social Media.

Hinkelmann, K., Ahmed, S. & Corradini, F. (2019). Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks - A Survey. AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering.

Horne, B. & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. Proceedings of the International AAAI Conference on Web and Social Media.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition.

Iol (2021). Independent Media demands health department come clean about Tembisa decuplets, stands by Piet Rampedi. [online] Available at: <https://www.iol.co.za/news/south-africa/gauteng/independent-media-demands-health-department-come-clean-about-tembisa-decuplets-stands-by-piet-rampedi-c3dbed32-a1ae-40cf-8168-0cd9a063ec9c> [Accessed 23 June 2021].

Ito, J., Song, J., Toda, H., Koike, Y. & Oyama, S. (2015). Assessment of tweet credibility with LDA features. Proceedings of the 24th International Conference on World Wide Web.

Jacovi, A., Shalom, O. S. & Goldberg, Y. (2018). "Understanding convolutional neural networks for text classification." arXiv preprint arXiv:1809.08037.

Jin, Z., Cao, J., Zhang, Y., Zhou, J. & Tian, Q. (2016). "Novel visual and statistical image features for microblogs news verification." IEEE transactions on multimedia 19(3): 598-608.

Kaliyar, R. K., Goswami, A. & Narang, P. (2019). Multiclass fake news detection using ensemble machine learning. 2019 IEEE 9th International Conference on Advanced Computing (IACC), IEEE.

Kaur, S., Kumar, P. & Kumaraguru, P. (2020). "Automating fake news detection system using multi-level voting model." Soft Computing 24(12): 9049-9069.

Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A. & Afroz, S. (2019). "A benchmark study on machine learning methods for fake news detection." arXiv preprint arXiv:1905.04749.

Klyuev, V. (2018). Fake news filtering: Semantic approaches. 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE.

Kochkina, E., Liakata, M. & Zubiaga, A. (2018). "All-in-one: Multi-task learning for rumour verification." arXiv preprint arXiv:1806.03713.

Koskinen, I., Zimmerman, J., Binder, T., Redstrom, J. & Wensveen, S. (2011). Design research through practice: From the lab, field, and showroom, Elsevier.

Kowsari, K., Heidarysafa, M., Brown, D. E., Meimandi, K. J. & Barnes, L. E. (2018). Rmdl: Random multimodel deep learning for classification. Proceedings of the 2nd International Conference on Information System and Data Mining.

LeCun, Y., Kavukcuoglu, K. & Farabet, C. (2010). Convolutional networks and applications in vision. Proceedings of 2010 IEEE international symposium on circuits and systems, IEEE.

Li, S., Ma, K., Niu, X., Wang, Y., Ji, K., Yu, Z. & Chen, Z. (2019). Stacking-Based Ensemble Learning on Low Dimensional Features for Fake News Detection. 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), IEEE.

LibGuides (2021). LibGuides: Critical Thinking and the News: News in the Age of Clickbait. [online] Available at: <https://fitnyc.libguides.com/c.php?g=631295&p=4411315> [Accessed 24 June 2021].

Liu, H. (2019). A Location Independent Machine Learning Approach for Early Fake News Detection. 2019 IEEE International Conference on Big Data (Big Data), IEEE.

Looijenga, M. S. (2018). The Detection of Fake Messages using Machine Learning, University of Twente.

Lu, Y. J. & Li, C. T. (2020). "GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media." arXiv preprint arXiv:2004.11648.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K. F. & Cha, M. (2016). "Detecting rumors from microblogs with recurrent neural networks."

Mahabub, A. (2020). "A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers." SN Applied Sciences 2(4): 1-9.

Mahwah, NJ., Riedel, B., Augenstein, I., Spithourakis, G. P. & Riedel, S. (2017). "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task." arXiv preprint arXiv:1707.03264.

Mandical, R. R., Mamatha, N., Shivakumar, N., Monica, R. & Krishna, A. (2020). Identification of Fake News Using Machine Learning. 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), IEEE.

Mayer, J., (2021). How Russia Helped Swing the Election for Trump. [online] The New Yorker. Available at: <https://www.newyorker.com/magazine/2018/10/01/how-russia-helped-to-swing-the-election-for-trump> [Accessed 12 May 2021].

Medium (2018). Why and how to Cross Validate a Model? [online] Available at: <https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f> [Accessed 12 July 2021].

Mehta, A. (2020). Everything You Need to Know About Linear Discriminant Analysis. [online] Available at: <https://www.digitalvidya.com/blog/linear-discriminant-analysis/> [Accessed 14 July 2021].

Mohale, P. & Leung, W. S. (2019). Fake News Detection Using Ensemble Machine Learning. European Conference on Cyber Warfare and Security, Academic Conferences International Limited.

Nickel, M., Murphy, K., Tresp, V. & Gabrilovich, E. (2015). "A review of relational machine learning for knowledge graphs." Proceedings of the IEEE 104(1): 11-33.

Okoro, E., Abara, B., Umagba, A., Ajonye, A. & Isa, Z. (2018). "A hybrid approach to fake news detection on social media." Nigerian Journal of Technology 37(2): 454-462.

Onan, A., Korukoğlu, S. & Bulut, H. (2016). "Ensemble of keyword extraction methods and classifiers in text classification." Expert Systems with Applications 57: 232-247.

Oshikawa, R., Qian, J. & Wang, W. Y. (2018). "A survey on natural language processing for fake news detection." arXiv preprint arXiv:1811.00770.

Pavlopoulos, J., Malakasiotis, P. & Androutsopoulos, I. (2017). "Deep learning for user comment moderation." arXiv preprint arXiv:1705.09993.

Peffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S. (2007). "A design science research methodology for information systems research." Journal of management information systems 24(3): 45-77.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J. & Fugelsang, J. A. (2015). "On the reception and detection of pseudo-profound bullshit." Judgment and Decision making 10(6): 549-563.

Pennycook, G. & Rand, D. G. (2019). "Fighting misinformation on social media using crowdsourced judgments of news source quality." Proceedings of the National Academy of Sciences 116(7): 2521-2526.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A. & Mihalcea, R. (2017). "Automatic detection of fake news." arXiv preprint arXiv:1708.07104.

Pérez-Rosas, V. & Mihalcea, R. (2015). Experiments in open domain deception detection. Proceedings of the 2015 conference on empirical methods in natural language processing.

Popat, K., Mukherjee,S., Strötgen, J. & Weikum, G. (2017). Where the truth lies: Explaining the credibility of emerging claims on the web and social media. Proceedings of the 26th International Conference on World Wide Web Companion.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J. & Stein, B. (2017). "A stylometric inquiry into hyperpartisan and fake news." arXiv preprint arXiv:1702.05638.

Qiao, Y., Wiechmann, D. & Kerz, E. (2020). A Language-Based Approach to Fake News Detection Through Interpretable Features and BRNN. Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM).

Qin, Y., Wurzer, D., Lavrenko, V. & Tang, C. (2016). "Spotting rumors via novelty detection." arXiv preprint arXiv:1611.06322.

Rampersad, G. & Althiyabi, T. (2020). "Fake news: Acceptance by demographics and culture on social media." Journal of Information Technology & Politics 17(1): 1-11.

Reddy, H., Raj, N., Gala, M. & Basava, A. (2020). "Text-mining-based fake news detection using ensemble methods." International Journal of Automation and Computing: 1-12.

Reddy, P. B. P., Reddy, M. P. K., Reddy G. V. M., & Mehata, K. (2019). Fake data analysis and detection using ensembled hybrid algorithm. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), IEEE.

Richey, R. C. & Klein, J. D. (2007). "Design and development research." Routledge.

Rosenkilde, T. C. (2017). "A Benchmark for Automated Fact Checking with Knowledge Bases." ASU Library.

Roy, A., Basak, K., Ekbal, A. & Bhattacharyya, P. (2018). "A deep ensemble framework for fake news detection and classification." arXiv preprint arXiv:1811.04670.

Rubin, V. L., Chen, Y. & Conroy, N. K. (2015). "Deception detection for news: three types of fakes." Proceedings of the Association for Information Science and Technology 52(1): 1-4.

Rubin, V. L., Conroy, N., Chen, Y. & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. Proceedings of the second workshop on computational approaches to deception detection.

Ruchansky, N., Seo, S. & Liu, Y. (2017). Csi: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.
Saeed, R. M., Rady, S. & Gharib, T. F. (2019). "An ensemble approach for spam detection in Arabic opinion texts." Journal of King Saud University-Computer and Information Sciences.

Sadiq, A., (2018). Re: What is the equation for the Area under the ROC curve?. Retrieved from: https://www.researchgate.net/post/What-is-the-equation-for-the-Area-under-the-ROC-curve/5ae8bac1cbdfd41ea93376b0/citation/download

Saeed, U., Fahim, H. & Shirazi, F. (2020). "Profiling Fake News Spreaders on Twitter." Notebook for PAN at CLEF.

Sagi, O. & Rokach, L. (2018). "Ensemble learning: A survey." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4): e1249.

Sangamnerkar, S., Srinivasan, R., Christhuraj, M. & Sukumaran, R. (2020). An Ensemble Technique to Detect Fabricated News Article Using Machine Learning and Natural Language Processing Techniques. 2020 International Conference for Emerging Technology (INCET), IEEE.

Science News for Students (2021). How to fight online hate before it leads to violence. [online] Available at: <https://www.sciencenewsforstudents.org/article/online-hate-violence-bigotry-racism-solutions> [Accessed 25 June 2021].

Scikit-learn (2021). scikit-learn 0.24.1 documentation. [online] Available at: <https://scikit-learn.org/stable/modules/sgd.html/> [Accessed 14 July 2021].

Shu, K., Mahudeswaran, D., Wang, S. & Liu, H. (2020). Hierarchical propagation networks for fake news detection: Investigation and exploitation. Proceedings of the International AAAI Conference on Web and Social Media.

Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. (2017). "Fake news detection on social media: A data mining perspective." ACM SIGKDD explorations newsletter 19(1): 22-36.

Shukla, Y., Yadav, N. & Hari, A. (2019). "A Unique Approach for Detection of Fake News using Machine Learning." International Journal for Research in Appplied Science & Engineering Technology (IJRASET) 7(VI).

Silva, R. M., Santos, R. L., Almeida,T. A. & Pardo,T. A. (2020). "Towards automatically filtering fake news in Portuguese." Expert Systems with Applications 146: 113199.

Sisodia, D. S. (2019). "Ensemble Learning Approach for Clickbait Detection Using Article Headline Features." Informing Sci. Int. J. an Emerg. Transdiscipl. 22: 31-44.

Smitha, N. & Bharath, R. (2020). Performance Comparison of Machine Learning Classifiers for Fake News Detection. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE.

Song, C., Yang, C., Chen, H., Tu, C., Liu, Z. & Sun, M. (2019). "CED: Credible early detection of social media rumors." IEEE Transactions on Knowledge and Data Engineering.

Sparrow, A., (2017). WhatsApp must be accessible to authorities, says Amber Rudd. [online] Available at: <https://www.theguardian.com/technology/2017/mar/26/intelligence-services-access-whatsapp-amber-rudd-westminster-attack-encrypted-messaging> [Accessed 12 May 2021].

Srivastavam,T. (2019). Evaluation Metrics Machine Learning. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/> [Accessed 14 July 2021].

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S. & de Alfaro, L. (2017). "Some like it hoax: Automated fake news detection in social networks." arXiv preprint arXiv:1704.07506.

Tammam, A. G., Sucipto, S. & Indriati, R. (2018). "Hoax Detection at Social Media With Text Mining Clarification SystemBased." JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika) 3(2): 94-100.

Tandoc Jr, E. C., Lim, Z. W. & Ling, R. (2018). "Defining “fake news” A typology of scholarly definitions." Digital journalism 6(2): 137-153.

Thota, A., Tilak, P., Ahluwalia, S. & Lohia, N. (2018). "Fake news detection: a deep learning approach." SMU Data Science Review 1(3): 10.

Ünal, R. & Çiçeklioğlu, A. Ş. (2019). "The Function and Importance of Fact-Checking Organizations in the Era of Fake News: Teyit. Org, an Example from Turkey." Media Studies 10(19): 140-160.

van der Linden, S., Roozenbeek, J. & Compton, J. (2020). "Inoculating Against Fake News About COVID-19." Frontiers in Psychology 11: 2928.

Vijayaraghavan, S., Wang, Y., Guo, Z., Voong, J., Xu, W., Nasseri, A., Cai, J., Li, L., Vuong, K. & Wadhwa E. (2020). "Fake News Detection with Different Models." arXiv preprint arXiv:2003.04978.

Viviani, M. & Pasi, G. (2017). "Credibility in social media: opinions, news, and health information—a survey." Wiley interdisciplinary reviews: Data mining and knowledge discovery 7(5): e1209.

Vlachos, A. & Riedel, S. (2014). Fact checking: Task definition and dataset construction. Proceedings of the ACL 2014 workshop on language technologies and computational social science.

Volkova, S., Shaffer, K., Jang, J. Y. & Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Waikhom, L. & Goswami, R. S. (2019). Fake news detection using machine learning. Proceedings of International Conference on Advancements in Computing & Management (ICACM).

Wang, W. Y. (2017). "" liar, liar pants on fire": A new benchmark dataset for fake news detection." arXiv preprint arXiv:1705.00648.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L. & Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining.

Wardle, C. & Derakhshan, H. (2017). "Information disorder: Toward an interdisciplinary framework for research and policy making." Council of Europe report 27: 1-107.

Webwise (2021). Webwise - Internet Safety. [online] Available at: <https://www.webwise.ie/> [Accessed 9 May 2021].

Wu, K., S. Yang & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. 2015 IEEE 31st international conference on data engineering, IEEE.

Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z. & Yu, P. S. (2018). "TI-CNN: Convolutional neural networks for fake news detection." arXiv preprint arXiv:1806.00749.

Your Dictionary (2021). Examples of Propaganda Done with Different Tactics. [online] Available at: <https://examples.yourdictionary.com/examples-of-propaganda.html> [Accessed 23 June 2021].

Yu, F., Liu, Q., Wu, S., Wang, L. & Tan, T. (2017). A Convolutional Approach for Misinformation Identification. IJCAI.

Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtelris, N., Leontiadis, I., Sirivianos, M., Stringhini, G. & Blackburn, J. (2017). The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. Proceedings of the 2017 internet measurement conference.

Zannettou, S., ElSherief, M., Belding, E., Nilizadeh, S. & Stringhini, G. (2020). Measuring and characterizing hate speech on news websites. 12th ACM Conference on Web Science.

Zannettou, S., Sirivianos, M., Blackburn, J. & Kourtellis, N. (2019). "The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans." Journal of Data and Information Quality (JDIQ) 11(3): 1-37.

Zeng, L., Starbird, K. & Spiro, E. (2016). # unconfirmed: Classifying rumor stance in crisis-related social media messages. Proceedings of the International AAAI Conference on Web and Social Media.

Zhang, J., Li, Z., Nai, K., Gu, Y. & Sallam, A. (2019). "DELR: A double-level ensemble learning method for unsupervised anomaly detection." Knowledge-Based Systems 181: 104783.

Zhao, Z., Resnick, P. & Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. Proceedings of the 24th international conference on world wide web.

Zhou, X., Jain, A., Phoha, V. V. & Zafarani, R. (2020). "Fake news early detection: A theory-driven model." Digital Threats: Research and Practice 1(2): 1-25.

Zhou, X., Wu, J. & Zafarani, R. (2020). SAFE: Similarity-Aware Multi-modal Fake News Detection, Cham, Springer International Publishing.

Zhou, X. & Zafarani, R. (2020). "A survey of fake news: Fundamental theories, detection methods, and opportunities." ACM Computing Surveys (CSUR) 53(5): 1-40.

Zhou, X., Zafarani, R., Shu, K. & Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. Proceedings of the twelfth ACM international conference on web search and data mining.

Zhubarb (2018). Area under Precision-Recall Curve (AUC of PR-curve) and Average Precision. [online] Available at: <https://stats.stackexchange.com/q/15701> [Accessed 23 June 2021].

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M. & Procter, R. (2018). "Detection and resolution of rumours in social media: A survey." ACM Computing Surveys (CSUR) 51(2): 1-36.

Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., Lukasik, M., Bontcheva, K., Cohn, T. & Augenstein, I. (2018). "Discourse-aware rumour stance classification in social media using sequential classifiers." Information Processing & Management 54(2): 273-290.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G. & Tolmie, P. (2016). "Analysing how people orient to and spread rumours in social media by looking at conversational threads." PloS one 11(3): e0150989.