Corroborating Information from Disagreeing Views Using Machine Learning Techniques

By

Tayyeba Riaz

00000277290

Supervisor

Dr. Seemab Latif

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Information Technology (MS-IT)

In

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST), Islamabad,
Pakistan.
(July 2022)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Corroborating information from Disagreeing views using Machine Learning Techniques" written by TAYYEBA RIAZ, (Registration No 00000277290), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: ______________________

Name of Advisor: Dr. Seemab Latif

Date: 30-Jul-2022

HoD/Associate Dean: ______________________

Date: ______________________

Signature (Dean/Principal): ______________________

Date: ______________________

Approval

It is certified that the contents and form of the thesis entitled "Corroborating information from Disagreeing views using Machine Learning Techniques" submitted by TAYYEBA RIAZ have been found satisfactory for the requirement of the degree

Advisor : Dr. Seemab Latif

Signature: ____________________

Date: 30-Jul-2022

Committee Member 1:Asad Ali Shah

Signature: ____________________

Date: 31-Jul-2022

Committee Member 2:Dr. Sharifullah Khan TI

Signature: ____________________

Date: 01-Aug-2022

Committee Member 3:Dr. Rafia Mumtaz

Signature: ____________________

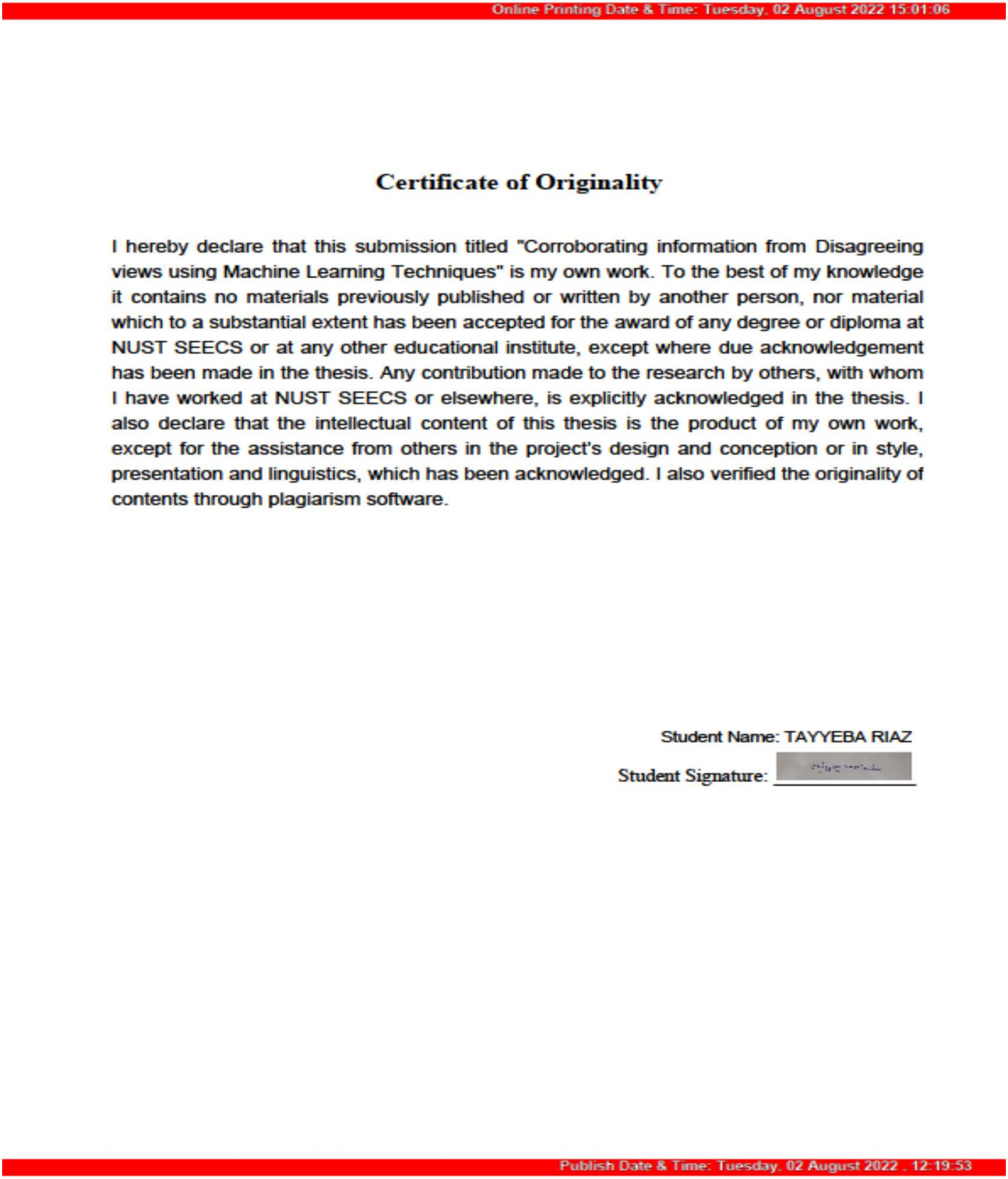
Date: 01-Aug-2022

Dedication

Dedicated to my parents,

Muhammad Riaz Akhtar and Farhat Parveen Hashmi,

who put their trust in me



Online Printing Date & Time: Tuesday, 02 August 2022 15:01:06

Certificate of Originality

I hereby declare that this submission titled "Corroborating information from Disagreeing views using Machine Learning Techniques" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: TAYYEBA RIAZ

Student Signature: ______________

Publish Date & Time: Tuesday, 02 August 2022 , 12:19:53

Acknowledgements

I own my deepest gratitude to Allah Almighty for His abundant blessings. I am sincerely grateful to my supervisor Dr. Seemab Latif for her continuous guidance and support. I would also like to thank my guidance and evaluation committee members Dr. Sharifullah Khan, Dr. Asad Ali Shah and Dr. Rafia Mumtaz for their valuable suggestions which helped me improve my work. I would also like to thank all my friends Najam-ul-Sahar, Aneeza Khalid, Abeer Gauher and Alia Usman for their suggestions, encouragement, and continuous support.

Finally, I would like to say a special thanks to my husband, Muhammad Azeem and my family for their endless love, kindness, and encouragement. Without their prayers and continuous support, it would not have been possible to accomplish this milestone.

Table of Contents

List of Abbreviations

Abbreviations	Full form
WWW	World Wide Web
MV	Majority Voting
CP	Credibility Perception
NA	Not Applicable
HAT	High Aesthetic Treatment
LAT	Low Aesthetic Treatment
GCN	Graph Convolutional Network
SIF	Smooth Inverse Frequency
LD	Levenshtein distance
ETBoot	Estimating Truth via Bootstrapping
vMF	von Mises-Fisher distribution
PAPRIKA	Potentially All Pairwise RanKings of all possible Alternatives
CC	Correlation Coefficient
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
WebCAST	Web Credibility Assessment Support Tool
NLP	Natural Language Pre-processing
ML	Machine Learning
TP/TN	True Positive/True Negative
FP/FN	False Positive/False Negative
SVM	Support Vector Machine

SGD	Stochastic Gradient Descent
NN	Neural Network
ROC curve	Receiver Operating Characteristic curve

List of Tables

List of Figures

Abstract

In this era of big data, huge amount of heterogeneous data is produced and shared on the internet making it a central medium for valuable sources of information. This data on the web can be published without quality control unlike the traditional media, thus, making it less reliable. Often data provided by different sources can be conflicting which can be due to noisy, erroneous, or obsolete data providers. It can also be easily manipulated by bots creating misleading data. This gives rise to a fundamental challenge for data extraction and fusion. This paper proposes an automated solution for truth finding from conflicting data by different sources by considering website credibility. It takes into consideration that different sources have varying degrees of reliability. It not only considers several factors about the sources but also provides with the true answer from a credible source. This paper identified seven web credibility categories namely Accuracy, Authority, Aesthetics, Professionalism, Popularity, Currency and Quality. Each category has several factors contributing to it. A total of 24 factors were used after applying feature reduction to approx. 100 identified factors from research. Six different supervised learning classifiers: Naïve Bayes, Support Vector Machine, Stochastic Gradient Descent, Neural Network, Decision Trees and Random Forest were employed. Existing solutions focus primarily on finding relevant web pages but either do not evaluate web pages' credibility rather focus on trustworthiness only or evaluate two to three out of seven credibility categories. Experiments on the Book-Author dataset shows that Random Forest performs the best with an accuracy of 97.45%, Precision 0.975, Recall 0.975 and F-measure 0.974 when all the categories are used collectively. This is significantly higher than the baseline method using a single factor that can be categorized to authority category. The baseline accuracy is 87.77% with a Bayesian based approach. Furthermore, different experiments using each category separately and in combination were performed which shows that categories with many factors contribute more to credibility than the ones with a single factor. These are Professionalism, Popularity and Quality. Also, the importance of aesthetics category is proved experimentally. Accuracy of 93.47% for aesthetics category alone shows that it is vital in credibility which is rarely recognized. However, this study focuses primarily on using all the seven categories for web credibility to resolve conflicting data.

CHAPTER 1

Introduction

1. Introduction

This chapter covers the introduction of this research. It starts with the background knowledge about the use of internet in the big data era, shares some related statistical information followed by the motivation of this research. Further, it describes the gaining need of the fourth V of Big data; 'veracity' and 'credibility' of websites. A brief description of limitations of current solutions leads to the definition of problem statement. It further contains research questions, research objectives and ends with explaining the thesis structure.

1.1. Background

Internet was initially invented for official usage for research, communication and connections [1]. In the twenty-first century, the internet has restructured the world around it. Access to information, communication and conducting commerce has been overpowered by the internet and it has changed ways people do it [2]. The internet usage has multiplied ten folds over time with varying scope. Indeed, the internet has revolutionised the world [1].

According to the Internet World Stats, there are 5.3 billion active internet users in the world [3]. The internet usage has increased by more than 900% in the world since 2005 [1]. In the start phase of Google in 1998, it was processing 10,000 search queries per day. This same amount was served in a single second by the end of 2006. Now, this number has transformed to over 40,000 search queries every second on average which translates to over **3.5 billion searches per day** and **1.2 trillion searches per year** worldwide [4]. Some of the commonly used categories of topics searched on Google in 2020-2021 were "news", "athletes", "how to help", "how to make", "recipes", "near me", "virtual", "where is..", "why", "during corona virus" [5].

In this era of big data, huge amount of data is produced and shared over different web sources increasing our dependency of daily life tasks on the use of the World Wide Web [6]. The amount of data on the internet is huge and of heterogeneous nature which creates many new challenges for information retrieval [7]. The web became the central medium for valuable sources of in-formation such as articles, blogs, and wikis, where people constantly share knowledge, report scientific studies, upload comments, and write reviews. As a consequence, the web emerged as the prime source for extracting and fusing information [8].

1.2. Motivation

Unlike the traditional media with good quality control, anyone can publish information on the World Wide Web. It is fast with no publishing costs, however, compromises the guarantee for the correctness of information. One of the fundamental difficulties of information extraction is that it can be erroneous, noisy, not up-to-date, biased and incorrect [9]. To add to the problem, different websites often provide conflicting information on a topic [6][10]. Such conflicting information can arise from disagreement, obsolete data or simply errors [11]. Often bots are used to exploit information and the website content can easily be manipulated [7].

A normal user with little knowledge would be unable to distinguish between a reliable source of information (e.g. a government website) vs a less reliable source (e.g. a personal blog). Even an expert would have to go through multiple sources for content verification which can be very time consuming. Also, returning incorrect and unreliable data in a query result can be misleading and can lead to harmful events [12]

A few examples of conflicting information:

Example 1 - Query on Ask.com: Height of Mount Everest [6]

Results: 29,035 feet (4 websites)

29,028 feet (5 websites)

29,002 feet (1 website)

29,017 feet (1 website)

Example 2 - Two different sources report the birth date of Napoleon Bonaparte as August 15, 1769 or as January 7, 1768. Also, many historians argue whether Napoleon was born French or Italian, [11]

Example 3 - Cast of the film 'Broken', [13]

Source 1 - Charlie Booty, Lily James, Tim Roth

Source 2 – Charlie Booty, Lily James

According to a survey on the credibility of websites conducted by Princeton Survey Research in 2005, 54% of people find news websites as trustworthy, 26% find commerce websites and 12% people find blogs trustworthy majority of the time [6][11].

Hence, this situation gives rise to the problem of *data fusion.*

1.3. Data Fusion and the fourth V of big data: Veracity

Data fusion is defined as “discovering true values from conflicting multi-sourced data” and corroborating it where corroboration means evidence supporting a statement [14].

Data fusion is a form of data integration where data from different sources like websites, social media, blogs etc. are combined. Such data is unreliable, has a lot of uncertainties and its sources have different degrees of accuracy [15]. Data integration has three broad goals: increasing the completeness, conciseness, and correctness of data. The main goal of data fusion, which is a part of this research, is to discover the correct data among the uncertain and possibly conflicting mined data [12].

With the advent of Big Data, quality of data and its source trustworthiness have become more important. In addition to the three salient features of big data, volume, velocity and variety, the fourth V of Big Data i.e. veracity is gaining more recognition in this era.

According to Li et al. [10], “knowing the precise trustworthiness of sources can fix nearly half of the mistakes in the best fusion results”

Thus, the data quality can be improved by data corroboration. It aims at resolving conflicts and finding the true values in a wide range of domains, including source selection in semantic web, semantic annotation cleaning in social networks, and information extraction [11]

1.4. Trustworthiness vs credibility

Trustworthiness and credibility are often used in synonymous manner. However, trust indicates dependability but trust in information is indicated by credibility.

Credibility can be defined as believability that results from evaluating multiple dimensions simultaneously. A vast majority of researchers identify “trustworthiness” and “expertise” as the two key components or dimensions of credibility.

Trustworthiness is defined as truthful, and unbiased (perceived goodness of the source) whereas Expertise is defined as experienced, and competent (perceived knowledge of the source). To evaluate an overall credibility, both the trustworthiness and expertise need to be assessed [16]. To be more accurate, a lot of more dimensions in addition to these two contribute to actual credibility of a source as per research.

Thus, S. Aggarwal et al. in [17] define credibility as “the level of confidence a user puts on the information available on a given website based on various objective and subjective factors”.

Credibility can comprise of multiple concepts or dimensions such as trust, reliability, accuracy, reputation, quality, authority and competence, and more where each concept may add up to trustworthiness or expertise; the two key components of credibility. Each category consists of several factors which collectively compute the credibility of a source [18].

1.5. Limitations of the previous research methods

Firstly, some of the previous suggested methods in this field are based on Majority Voting (MV) i.e. count the number of occurrences of each answer. But many bots or sources often replicate and maliciously spread false information. In the *example 1* given above, none of the provided answers including the majority provided answer is correct. The correct answer is 29029 feet.[19]

Secondly, researches have been conducted on evaluating websites according to *their authority (qualified enough, has some knowledge and respect), or popularity (more visits, reads and shares)* but these alone do not guarantee accuracy of information.

For example, Google ranked two bookstores (Barnes & Noble and Powell’s books) on top with many errors on authors’ information while another small bookstore (A1 Books) provided more correct information [6].

Different sources on the web have different qualities which also vary over time. Also, the credibility of a source is not known apriori [20]. Thus, this shows that the quality of the information and source credibility should be considered when corroborating answers to identify the correct answer to a query [11]. More limitations based on researches are discussed in section 2.4.

Generally in truth finding methods, if a credible source provides data, it is considered to be true. Likewise, the source is credible if it provides true data [6]. Thus, both data and source can be checked to measure their *credibility*. The credibility of a website depends on its facts' accuracy rather than the number of facts it provides [20].

Therefore, as mentioned in section 1.4., it is vital to propose an automated solution for truth finding that not only considers several factors about the sources but also provides with the true answer from a credible source.

This study is aimed to be a subsequent contribution in an attempt to provide an automated solution to the aforementioned limitations. Hence, the most significant challenge for this task is to estimate source credibility and select the answers supported by high quality sources to resolve conflicts.

1.6. Problem Statement

"To infer the veracity of online (conflicting) information being claimed by different sources on the web and identifying the true value from a credible source."

1.7. Research Questions

Credibility of web information is highly important to avoid misleading or harmful information in every field. To measure the credibility, it is important to find the right set of website (source) factors and their categories. Also, to examine the factors and their categories used by state of the art methods and whether they have used all the categories of web credibility as defined above.

Motivation of the research encourages us to answer the following research questions:

RQ1. What credibility categories are required to be considered to evaluate websites' credibility?

RQ2. What percentage does each category contribute to computing credibility score?

RQ3. What features under each category can be used to measure website credibility for trustworthy data?

RQ4. Does the visual appearance (aesthetics) play a role in website credibility? If so, how much?

1.8. Research Objectives

Research objectives highlights the necessary steps that should be taken in order to find the answers discussed in research questions section. A survey of existing solutions has been completed to identify the factors/features used by other researchers to compute credibility of web information, their datasets and methodologies. The purpose of this research is to find the best suited and contribution of each category and its relevant features (factors) in web information credibility in order to resolve conflicts in answers. This study proposes the following research objectives to satisfy the research questions:

RO1. To identify the best suited categories for information credibility

RO2. To identify the contribution of each category in credibility

RO3. To identify the best suited factors in corresponding categories for credibility.

RO4. To identify the role and need of aesthetics in web credibility

1.9. Thesis Structure

The structure of this thesis research is as follows: Chapter 2, the literature review, provides a survey of data fusion and web credibility researches done from 1999 to 2020. It also identifies the factors used by other researchers to achieve credibility and resolving conflicting data. In addition, it identifies the research gap between state of the art methods and the proposed solution in regards to the credibility categories. This is followed by chapter 3, methodology in which the implementation of the proposed method is discussed. The system is evaluated against multiple machine learning classifiers using stratified cross and hold out validation. Classifier measures, performance and each credibility category evaluation and their results are discussed in chapter 4. At the end, this study concludes the findings and future work in chapter 5.

CHAPTER 2

Literature Review

2. Literature Review

It is a vital and a challenging task to find out the most likely true claim from multiple conflicting values provided by multiple sources in truth discovery process. Many traditional and inexperienced methods, such as voting, do not pay attention to the web source's reliability, and hence may fail in particular cases. These methods do not help the users build a trust relationship between them and the information source based on the quality of information and the source. Therefore, many new researches propose methods to evaluate web source trustworthiness in an accurate manner as discussed below [21][22].

The factors to evaluate credibility of a source can be categorized into seven categories i.e. accuracy, authority, professionalism, popularity, currency, impartiality and quality. This is based on the research of authors in [18] who reviewed and summarized many existing researches from 1996 onwards. Existing solutions focus primarily on finding relevant web pages but either do not evaluate Web pages' credibility rather focus on trustworthiness only or evaluate two to three out of seven credibility categories. Each category of information credibility identification is described in Table 2.1. Each of these categories focuses on one aspect of information credibility. To measure the credibility of a source for a true value, each of these should be considered and collectively used to attain a good accuracy score [18].

***Table 2.1*. Description of credibility categories**

Category name	Explanation
Accuracy	Correctness of website content
Authority	Experience and popularity of the website
Professionalism	Efficiency of a website with its policies and features available on the website
Popularity	Website's reputation among web users and reviewers
Currency	Frequency of update of website content
Impartiality	Biasness of the content
Quality	Rating of the information in terms of its readability

Most of the current methods use probabilistic models to iteratively compute and update source trustworthiness and confidence of their claims and their weights respectively. They work with a principle that a source is higher in weight if it provides true values often and is more trustworthy with more contribution in the truth estimation step. Thus, weights play a vital role in truth estimation [23].

The following algorithm flowchart represents the basic principle that most of the current approaches work on.

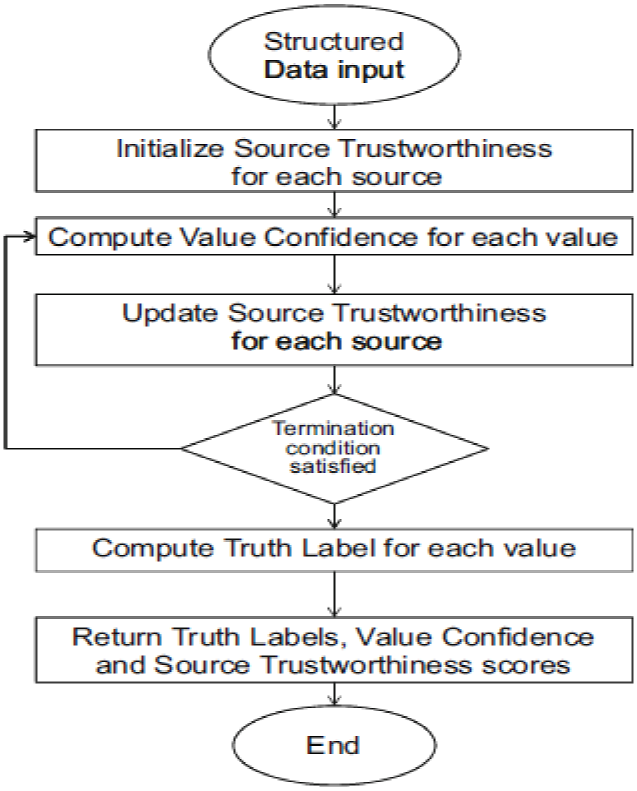

Structured Data input
Initialize Source Trustworthiness for each source
Compute Value Confidence for each value
Update Source Trustworthiness for each source
Termination condition satisfied
Compute Truth Label for each value
Return Truth Labels, Value Confidence and Source Trustworthiness scores
End

Figure 2.1 **Basic principle flowchart of current techniques (Image taken from [20])**

2.1. Categories of truth finding approaches

2.1.1. Traditional truth finding techniques

Some of the traditional methods that cannot be assigned to any category are discussed here first. One of the earliest and simplest traditional method to find true values is Majority Voting (MV). MV regards a value as true if claimed by majority of the sources. It randomly selects a value in case of a tie where the chance to deduce a wrong answer is 1/|V0a|. It is a simple method. MV assigns equal weights to each source and assumes that each source has the same quality. This is opposed to reality where sources differ in qualities, coverage and scope. Also, the properties of the sources and the claims is unknown.

Authors discuss a variant of MV called weighted voting-based methods which follow the principle that the information from reliable sources will be counted more in the aggregation. It considers ‘corroboration’ i.e. taking into account trust in the views [24].

Authors in this paper [25] further discuss two traditional methods; Average-Log and investment. These are described below.

The Average-Log method considers the confidence of claims and uses their Average-Log. It employs Sums’ Bi update rule to compute trustworthiness. It is a less complex method, however, if the number of claims is relatively low, it overestimates the source trustworthiness because of simply using average.

The Investment method proposes that trustworthiness of sources is “invested” uniformly between the claims. The confidence of claims increases non-linearly and the sum of these confidence of the claims adds up to a source’s trustworthiness which is weighted according to the trust ratio previously contributed to each (relative to the other investors). The higher confidence claims due to higher-trust sources are more believed and those sources are considered more trustworthy. This method is based on common sense and converges at a fast rate but it requires prior knowledge on the subject.

Table 2.2 Traditional truth finding techniques

	Research paper	**Methodology used**	**Results achieved**	**Limitations**
1	Majority Voting [24]	regards a value as true if claimed by majority of the sources	Simple, less complex	assigns equal weights to each source considering same trustworthiness
2	Average-Log [25]	Computes average Log of the confidence of claims and employs Sums’ Bi update rule to compute trustworthiness.	Less complex	Overestimates trustworthiness with low no. of claims, single credibility category

3	Investment [25]	Invest source trustworthiness in claims whose confidence adds upto source trustworthiness based on investment ratio	Simple, based on common sense and converges fast	requires prior knowledge on the subject

2.1.2. Recent truth finding techniques

2.1.2.1. Accuracy

Accuracy requires that a web source is correct and free of errors to a certain level and the information on the website is verifiable offline [26]. It deals with the correctness of information provided by the author [18].

The authors in [6] proposed a method called TRUTHFINDER, which claims that a dependency relationship exists between websites and facts, i.e., if a website provides many true values or facts, it is considered to be trustworthy. Likewise, if a fact is claimed by many trustworthy websites, it is likely to be true. They also claim that often facts can be slightly different yet may support each other. For example, two different websites claim that a certain camera is 4 inches and 10 cm long respectively. If one of these facts is true, it renders the other true automatically. They also consider influences between facts. A fact is likely to be wrong if it conflicts with another fact with high confidence provided by many trustworthy websites. Trustworthiness of websites and the correctness of information i.e. probabilities of facts being true or the fact confidence are derived from each other using an iterative computational method until it reaches a stable state. It achieved an accuracy of 0.95/95 percent in discovering true facts, and it can select better trustworthy websites than popularity-based search engines such as Google.

They solve the limitation of Authority-Hub analysis [14] . They claim that the number of facts provided by a website does not define its trustworthiness rather the accuracy of those facts does. This is in contrast to Authority-Hub analysis which computes trustworthiness of a website by adding up the weights of its facts.

The next study computes three measures: accuracy of sources, dependence between sources, and confidence of values. Their accuracy model is based on TRUTHFINDER but additionally considers dependence between sources in finding true values. The authors in [27] claim that a dependency relation exists between ‘sources’ and apply Bayesian analysis to iteratively compute the probability of two data sources being dependent. They claim that a source is dependent and copies from another when they share a large number of common values that are rarely provided by other sources (e.g., particular false values). It is not necessary that a complete dependency relation exists between two sources if they share the same true value, but a rare event of sharing the same false values shows full dependency between sources. The probability of providing a true value is the same for all independent sources for each object. This method requires identifying if dependency exists between two sources by sharing same values and also whether the common values are true or false to deduce the copier source. The accurate data sources i.e. providing true values have higher weights and thus are more trustworthy. This research is the first to analyse source dependence in truth discovery which significantly improved accuracy and is far more scalable with large number of sources.

Table 2.3 **Summary of accuracy category research papers**

	Research paper	Factor used	Factor category	Methodology used	Results achieved	Limitations
1	Truth Discovery with Multiple Conflicting Information Providers on the Web [6]	Correctness of information based on interdependency between sources and facts and influence between facts	Accuracy	Iterative computation of website trustworthiness and value confidence score (facts influence from each other) – Bayesian based	95% accuracy, better trustworthy websites than popularity based search engines	uniform initial source trustworthiness i.e 0.9, single credibility category
2	Integrating Conflicting Data: The Role of Source Dependence [27]	Correctness of information based on dependency between data sources	Accuracy	Iterative Bayesian analysis to find the copier source sharing large false values	Robust and effective in preventing falsification	uniform initial source trustworthiness, single credibility category

2.1.2.2. Authority

Authority deals with experience and popularity of the source providing the content [18]. To determine the authority of a website, information about its author like the author's qualifications and credentials in the web community and whether the site is recommended by a trusted other is important [26].

Authors in [14] propose a methodology to compute source trustworthiness based on in-degree hyperlinks (no. of pages that have links to a page) in the authority category. They refer the two set of pages as authorities and hubs. Hubs are the pages that have links to multiple relevant authoritative page while the authority pages is the initial result set of a query which has authority on a common topic. Thus, the hub pages which point to the authoritative pages have an overlap in them. Authorities are calculated as the sum of the scaled hub values that point to that page and hubs value is calculated as the sum of the scaled authority values of the pages it points to. These sums are called weights which are maintained and update via an iterative algorithm. The authors make use of directed graphs to form links between pages (nodes). A link from a page to another applies former's authority over the later. Most authoritative and thus trustworthy page has the greatest number of in-links. The authorities are found only through the pages pointing to them which focuses on the computational effort but has a drawback that many in-degree links are often created for directions or advertisements purposes. This does not guarantee authority alone.

The authors in this paper [28] resolve conflicts by incorporating the domain expertise knowledge/data richness of a website which can be mapped to the authority category. They claim that sources have different expertise in different domains and hence can have different reliabilities. They apply the Bayesian approach to derive domain expertise of each source from various domains by considering its information richness. The authors make use of a book-author dataset which shows that a bookseller can have more science category books data than data for arts category books making it information rich in the science category. They also study mutual influence between domains. Their method has an advantage of finding multiple possible truths in an unsupervised way. As opposed to single truth problems, they consider partially correct answers as supportive to full correct answers instead of totally ignoring them, thus, naturally supporting multiple truths for an object. This methodology performs effectively and efficiently with significantly reducing error rates.

This study [29] uses a probabilistic graph model to construct a relationship between sources, objects (questions) and truth. They have redesigned the metrics of source quality taken from the Authority Hub [14] method considering how source quality is affected when null is provided by the sources. Their algorithm significantly increases recall compared with the Authority Hub algorithm

Table 2.4 **Summary of authority category research papers**

	Research paper	**Factor used**	**Factor category**	**Methodology used**	**Results achieved**	**Limitations**
1	Authoritative Sources in a Hyperlinked Environment [14]	Indegree hyperlinks	Authority	Iterative computation of hubs and authority pages – Bayesian based	Low computational effort to find trustworthy website	Indegrees can be for directions or advertisements - not for support
2	Domain-Aware Multi-Truth Discovery from Conflicting Sources [28]	Domain expertise knowledge/data richness of a website	Authority	Bayesian approach to derive domain expertise of each source from various domains based on its information richness	Significantly reduces error rates. Achieved precision 0.91, recall 0.89, F1 measure 0.90	No significant increase in precision, uniform initial source trustworthiness, Data keeps updating so factor not always possible to compute, single credibility category
3	An effective truth discovery algorithm with multi-source sparse data [29]	how source quality is affected when null is provided by the sources.	Authority	Based on Authority-Hub analysis [14]	Significantly increases recall	Indegrees can be for directions or advertisements - not for support and hence can be biased

2.1.2.3. Aesthetics

Aesthetics of websites refers to the visual appearance of websites including fonts, colours, layouts, structure and presentation of data etc. [18]. According to some studies, aesthetics plays its role in changing people's perception about credibility of a website. Researchers claim that there is a correlation between aesthetics/design and credibility judgment [30]. However, aesthetics alone do not make credible web pages. But an aesthetically pleasing website increases its credibility in people's perception incrementally [31].

This paper [22] discusses the credibility factors on the web and online social media. Due to interface design differences on both platforms and scope limitation, the social media factors will not be discussed in this study. There are two systems to find credibility indicators namely heuristic and analytic. This paper mainly focuses on the heuristic measures to find credibility and claim that studies incorporating the reader's heuristic element (subjective factors) are a minority.

The authors in this paper claim that it is entirely dependent on the web readers to choose the process to judge the veracity of a piece of information with the sheer amount of online data available. This process is known as reader's credibility perception (CP). Factors such as author's location, reader's tech experience and verification steps influence the reader's CP. Also, a reader's demographic attributes and gender influence interface, expertise and security preferences. However, the main factor of interest of author's is the cognitive heuristics. It includes the interface design layout with less advertisements and better UI/UX. The time indicator to show recency of information. Such websites are found more credible.

The author's also discuss the important credibility factors for analytic system i.e. the source reputation and trustworthiness. Also, composition of information shows the level of relatedness to the topic and relevant to keyword search. A language element involving the use of sentiment and semantic words to describe the positive/negative relation of the post towards the topic. Embedding external sources in the information is an option for readers to do their own verification of the credibility of the information.

Authors in [32] studies the possible correlation between aesthetics or design of a website and users' credibility judgements. The study aimed at finding the instant response of users captured by first impressions which decides whether the user will stay or move on to the next site. It is referred to as amelioration effect of visual design and aesthetics on content credibility. The findings indicated that High Aesthetic Treatment (HAT) resulted in better ranking of website credibility by the users generally. However, the authors claims that aesthetics alone does not imply a credible website, but it increments the credibility level. Thus, the result of experiment might be due to the limitation that subjects were inexperienced and not knowledgeable about accessibility and usage of site. Similarly, they might be knowledgeable and would know the credible resources on the site. The research lacked conclusions on which features/elements of website design affected credibility exactly.

Table 2.5. **Summary of aesthetics category research papers**

	Research paper	**Factor used**	**Factor category**	**Methodology used**	**Results achieved**	**Limitations**
1	A Review on Credibility Perception of Online Information [22]	Cognitive heuristics - interface design layout with less advertisements and better UI/UX. The time indicator to show recency of information.	Aesthetics	A review of credibility factors influencing reader’s credibility perception	Many subjective factors summarized in the study	Not Applicable (NA)
2	Aesthetics and credibility in web site design [32]	Viewing of HAT vs LAT webpages	Aesthetics	Finding the instant response of users captured by first impressions of HAT vs LAT webpages	HAT resulted in better ranking of website credibility by the users	Subjects were either inexperienced or if knowledgeable knew the credible sources, aesthetics features affecting credibility unfound, single credibility category

2.1.2.4. Professionalism

Professionalism deals with meta-data about privacy policy, rating of a website, its loading speed, properly used spellings and grammar for the content and mobile friendliness etc. Security can be a sub-category of professionalism. It deals with factors like declaration of secured protocol utilization, approval of security tokens like TTP [33], presence of security policies [18], http/https, and presence of malicious ads.

There are a few studies that consider professionalism category for credibility among few others. As these are multiple credibility categories' studies, these are discussed in section 2.3.

2.1.2.5. Popularity

Website's reputation among web users and reviewers is its popularity [18]. Building websites in attractive and information-rich way attracts more users and pages leading to high link popularity. Thus, leading to more reliable values from central pages in a tremendous way [34].

The authors in [7] use backlinks factor to find reliable web sources. They invented a method, called PageRank, to compute a ranking for every web page. This method did not consider what content a webpage consists but it made use of the web page's location in the web's graph structure to rank the webpage. The now famous and most used web search engine, Google, was developed by the authors to test PageRank.

PageRank is based on peer review system where if the total sum of backlinks of a page is high, it has a high rank. Such a page may have a huge number of backlinks or few backlinks which are highly ranked and more significant. Once the rank is assigned, it is evenly divided among its forward links to contribute to the ranks of the pages they point to. This rank is lost for pages with no forward links. This method works recursively until convergence. However, ranks are accumulated without further distribution if two pages point to each other only with a web page pointing towards one of them. It creates a trap loop. PageRank provides high quality results with answers from most important, popular and central webpages.

Authors et al. in [13] claim that each object or question can have multiple true values. If values of an object are partially correct from a source, they overlap the completely correct values from another source for the same object. This creates an implicit support and endorsement relation between the two. Furthermore, a source endorsed by many sources is regarded more authoritative on its data and hence trustworthy. Similarly, other sources are also considered trustworthy if they are endorsed by these authoritative sources

From a set of answers for a given object, values claimed by a source as its answer are positive claims. The rest answers from the set are negative claims or disclaimed values for that source and object. They use ±Agreement Graph to model agreement between sources on their positive and negative claims respectively. Random walk computations are performed on both graphs referred as Markov chain to derive two sided vote counts of endorsement between sources and to finally evaluate value veracity. This method achieved a precision of 0.90, recall of 0.92 F1 score of 0.91 where the latter two are higher than other applied methods on the given dataset. Also, this method has better convergence rate and accuracy as it does not require uniform initial source trustworthiness but automatically derives it by capturing source endorsement relations without using any prior knowledge.

An extension of this paper is proposed in [35] where authors proposed a graph-based model, called SmartVote, as an advanced solution for truth finding with multiple true values. This model incorporates four important implications, including two types of source relations, object popularity, loose mutual exclusion, and long-tail phenomenon on source coverage, for better truth discovery.

In addition to the supporting relation between sources modelled in the previous version; SourceVote [13], they further extend and incorporate the copying relation i.e. If a lot of false values are common between two sources, they are likely to copy from each other.

***Table 2.6*. Summary of popularity category research papers**

	Research paper	**Factor used**	**Factor category**	**Methodology used**	**Results achieved**	**Limitations**
1	The PageRank Citation Ranking: Bringing Order to the Web [7]	No. of backlinks of a webpage	Popularity	Iterative computation to find rank of webpage using web’s graph structure and its further division to forward links, until convergence	Answers from high quality, central and popular webpages found.	Complex with trap loops, single credibility category
2	SourceVote: Fusing multi-valued data via inter-source agreements [13]	Implicit endorsement relation between two sources with values overlap	Popularity	Applied Markov chain random walk computations on ±Agreement Graph to derive endorsement relation among sources	Outperforms the baselines (best recall 0.92 and F1 score 0.91) no uniform initial source trustworthiness (better accuracy and convergence rate	no significant precision increase, single credibility category
3	SmartVote: a full-fledged graph-based model for multi-valued truth discovery [35]	Endorsement (supportive) and copying relations, object popularity	Popularity	As above	Increased precision and F1 score by considering long tail phenomenon	Decrease in recall

2.1.2.6. Currency

Currency deals with the recency of information and its frequency of update. More up-to-date websites provide better information. Researches in this category often deal with coverage which refers to the comprehensiveness or depth of the information provided [26].

In this paper [36], the authors study the problem of finding true values by determining the copying relationship between sources, when the update history of the sources is known. The study uses coverage, exactness and freshness factors over time in a probabilistic model to infer the quality or reliability or sources. A Hidden Markov Model determines if two sources have a copying relation and which source copies from the other. It further determines the specific moments at which the copier source copies. Then a Bayesian model is developed to determine the true value of an object by aggregating information from the sources, and the evolution of the true values over time. This study achieved accuracy of 95% with higher coverage but an average accuracy of 88%. However, this technique has high scalability.

Authors in [37] propose estimating accuracy of facts by incorporating dependency between sources based on copying relations. They adopted techniques presented in [36] to compute accuracy but their dependency model differs and is improved. Their copying detection techniques are proposed for global copying detection on static data and these techniques can be extended for dynamic data following the ideas in [36].

They consider that the copying relationships can be complex: some sources copy from or are copied by multiple sources on different subsets of data; some co-copy from the same source, and some transitively copy from another.

This framework works in two steps. The first locally determines the copying correlation and the copying direction between each pair of sources in isolation and the second determines co-copying and transitive copying globally using the right evidences in local detection in a probabilistic model. In the first step, different copying evidences are plugged in including common mistakes as important evidences. But other important evidences are neglected such as same data formatting, similar real world objects provided by two sources. Some possible copying correlations are also neglected. For example, a source that copies the name of a book tends to also copy its author list. Results showed that these limitations often lead to wrong copying directions which affect the global copying detection as well. This algorithm performs effectively and efficiently based on the experimental results.

Table 2.7 **Summary of currency category research papers**

	Research paper	**Factor used**	**Factor category**	**Methodology used**	**Results achieved**	**Limitations**
1	Truth Discovery and Copying Detection in a Dynamic World [36]	Coverage, exactness and freshness of data with simple copying relation between sources with unknown update history	Currency	Hidden Markov model to determine copying relation and its direction and the exact time of copying. Bayesian model to find true values.	Accuracy of 95% with higher coverage but an average accuracy of 88%. High scalability.	Single credibility category
2	Global detection of complex copying relations - DEPEN and ACCU [37]	Same factors as above with more complex copying relations	Currency	2-step detection of copying relations respectively. Local detection of simple copying relations and global detection of complex copying relations with plugging in evidences	Effective and efficient methods	Neglecting important evidences in local detection results in wrong copying direction affecting global copying detection, single credibility category

2.1.2.7. Quality

Quality is the rating of the information in terms of its readability. This study [11] proposes fix point computation techniques that derive estimates of the true value of facts provided by multiple sources by taking into account fact confidence, as well as estimates of the quality of the sources.

The authors introduce probabilistic model that consists of three algorithms; COSINE (uses cosine similarity measure), 2-estimates (uses 2 estimators for the truth of facts and the error of views that are proved to be perfect in some statistical sense. Initially, all parameters are set for the views to be true, Then a set of parameters are estimated at one time until convergence), 3-estimates (refined version of 2 estimates that also estimates hardness of facts) that estimate the truth values of facts and trust in sources based on the quality of data. They all refine these estimates iteratively until a fix-point is reached. Their baseline methods are Voting, Counting and TRUTHFINDER. The COSINE, 2-estimates and 3-estimates resulted in a global precision of 88.1%, 88.2% and 91.5% respectively which is higher than the baseline methods (84%). However, 2-estimates is often very unstable and may perform worse than the baselines for a large range of parameters and 3-estimates performs better taking hardness of facts into account.

The authors in [38] propose a framework to find correct answers by taking into account how distinctively answers for a given query are reported within the websites (sources). Additionally, the frequency of answers in the search engine result and the relevance and originality of the sources reporting answers for a given query are considered. Each of these individual answers are assigned a score based on the above aspects for the likelihood of it being correct. Out of these, the similar answers' scores are aggregated. This study extracts many queries from TREC Question Answering track and a log of real web search engines to perform experiments. The results suggest that extracting answers from web pages of good quality in the presence of low quality data in a corroborative way provides with correct answers for a majority of queries faster.

***Table 2.8*. Summary of quality category research papers**

	Research paper	**Factor used**	**Factor category**	**Methodology used**	**Results achieved**	**Limitations**
1	Corroborating information from disagreeing views [11]	Similarity between facts and sources, also considering hardness of facts	Quality	Probabilistic model to compute similarity between facts and sources, iteratively computing trust in views and facts while considering error in views and hardness of facts	The 3 algos achieved precision 88.1%, 88.2% and 91.5% respectively higher than the baseline (84%).	2-estimates is unstable with large no. of parameters, Single credibility category
2	A framework for corroborating answers from multiple web sources [38]	Distinction of answers within sources, frequency of answers in query result set, relevance and originality of sources	Quality	Use of Zipf law model to use low ranked pages to add upto high score in the probabilistic model	PerCorrect 0.8, MRR 1.0 and faster extraction of correct answer with high quality data	Single credibility category

2.2. Truth discovery from text-based unstructured data

This study [39] makes use of answer space mining and fusing the semantic information from unstructured noisy text data in a graph instead of evaluating source trustworthiness directly. Using Graph Convolutional Network (GCN), it inputs graph of answers and outputs the rank of answers based on the identified truth answer vector. Each answer provided by sources for a given query is transformed to a real valued vector via Smooth Inverse Frequency (SIF). An undirected graph of these vectors is constructed. Semantic meaning of the answers are infused by the layer wise convolution operation such that each answer can obtain semantic information from neighbours. It is input to the GCN where finally, the answer credibility can be learnt by neural network. It sums up all the feature vectors of all neighbouring answers to improve the accuracy and efficiency of truth discovery. Removal of outliers highly impacts the accuracy. However, if number of real answers is smaller than the noisy answers and noisy data, performance will be unsatisfied. A future proposal for the method improvement suggests using less number of parameters.

The technique proposed in [8] uses a probabilistic model to capture the relationships between data sources, their contents, and the underlying factual information to output credible claims with credibility precision guarantee. The dependent factors are automatically learnt in an iterative learning process for best performance without any training. Encoded with a set of features, sources and documents are infused in a graph network. The maximum factual information possible is instantiated by the best parameters upon convergence. Features can be content-based, such as semantic features (e.g. category, entities, keywords), sentiments features (e.g. subjectivity, linguistic characteristics of a document), and syntactic features (part-of-speech tag, punctuation marks, spelling errors), advertisements, and page layout. Features can also be network-based, such as the overall ratings of sources sharing the same claims. Features can be derived from, e.g., activity logs (e.g. number of documents posted, frequency of updates), and demographic information (e.g. age, gender). To make features space finite, approximations are made. With increased no of features, the precision increases while decreasing the output size. They reuse these features to reduce total cost of samples per iteration. The method achieved a precision of 0.92 and recall of 0.72 with 0.9 precision threshold, outperforming baselines up to 6 times better. It is robust with noisy data when it is small, however, the performance declines with large number amount of noisy data

Authors in [40] propose a probabilistic model named TextTruth that uses clustering to group similar semantic words (factors) of keywords extracted from the answers. The method uses vector representations of keywords as inputs, infers trustworthiness of each answer factor and the provider in a probabilistic model and outputs the ranking of answers based on the trustworthiness of key factors within each answer. It generates the mixture of factors according to the Dirichlet distribution and the keyword embedding vector via a von Mises-Fisher (vMF) distribution. Semantic meaning of answers maybe complicated, thus, as opposed to other methods, clustering keywords based on semantics in fine grained clusters (factors) allows to estimate the trustworthiness of each answer factor instead of the whole answer and infer the correctness of each factor in the answer. The trustworthiness of keywords within each cluster is almost the same as they share similar meanings. User reliabilities are also dependent on their answers keywords belonging to correct or incorrect clusters factors. The proposed model naturally supports the partial correct answers. Many previous methods rank answers only based on semantic similarity between the question and answer while the question does not cover all the semantics that should be in an ideal answer. Thus, only relevant answers are discovered instead of trustworthy answers, unlike this method. Experimental results prove the effectiveness of this model.

This paper [41] proposes an approach to identify conflicting data on the web, making use of deviation in the embedded structured data on the web. Pre-processing of data is done to make the sources comparable. The detection algorithm uses (a) Levenshtein distance (LD) to represent the degree of conflict between data elements, (b) cosine similarity between vectors of LD values, and (c) a novel concept of a user-configurable control parameter called sensitivity vectors which encodes specific kind of conflict characteristics subject to investigation at runtime and then ranks the conflicting data as output. Investigating various conflicting data characteristics provides flexibility in the approach. The model does not require any training or parameter estimation which has an added advantage

The study [23] proposes a method, Estimating Truth via Bootstrapping (ETBoot), that focuses on leveraging the properties of bootstrapping and illustrate the importance of confidence interval estimation instead of fixed point in truth discovery. The study focuses on eliminating two limitations of some existing truth discovery solutions.

First, most existing truth discovery methods directly apply weighted averaging or voting using all sources' information, so they are sensitive to outlying claims. In contrast, ETBoot first bootstraps multiple sets of sources and then on each set of the bootstrapped sources it obtains a truth estimate based weighted averaging for continuous data or weighted voting for categorical data. The final truth estimator is defined as the mean of these estimates. ETBoot is more robust to the outlying claims and can achieve a better estimate of the truth. Second, many existing methods focus on point estimation of the truth, where important confidence information is missing. ETBoot can also construct an alpha-level two-sided confidence interval of the estimated truth.

Table 2.9 **Summary of text based truth finding techniques**

	Research paper	**Feature used**	**Methodology used**	**Results achieved**	**Limitations**
1	An unsupervised approach of truth discovery from multisourced text data [39]	answer space mining and semantic information fusion for each answer in a graph	SIF for vector transformation of answers. GCN to learn credibility	GCN complex relationship between source and claims at runtime, reduces running time for not computing source trustworthiness, More accuracy without outliers	Unsatisfied performance with less real answers than noisy answers, dependent on a lot of parameters
2	Maximal fusion of facts on the web with credibility guarantee [8]	Learns factors automatically in learning process with underlying factual information	probabilistic model to capture source-content and its factual information relationship and graph network	achieved precision 0.92, recall 0.72 outperforms baselines up to 6 times better, precision increases with more features	Output size decreases with more features, the performance declines with large number amount of noisy data
3	TextTruth: An Unsupervised Approach to Discover Trustworthy Information from Multi-Sourced Text Data [40]	Based on trustworthiness of similar semantic words in each answer	Uses Dirichlet and vMF distribution for factors generation and keyword embedding vector, clustering to find similar semantic words and probabilistic model to output answer ranks	Discover relevant and trustworthy answers, effective method unlike previous	NA

4	A Flexible Algorithmic Approach for Identifying Conflicting/Deviating Data on the Web [41]	Deviation in the embedded structured data on the web	LD for degree of conflict in data, cosine similarity between vectors of LD values, and sensitivity vectors to encode conflict characteristics found at runtime then ranks the conflicting data as output	Investigating data characteristics at runtime provides flexibility, no training or parameter estimation required	NA
5	Towards Confidence Interval Estimation in Truth Discovery [23]	Weighted averaging or voting.	Weighted averages/voting with bootstrapping	Not sensitive to outlying claims due to bootstrapping, confidence interval estimation of truth	NA

2.3. Web credibility categories

Studies that included multiple credibility categories and their factors in their research are discussed in this section.

The authors in [26] suggest five important credibility criteria that should be included in a method to evaluate website credibility by users. These include checking the accuracy, authority, objectivity, currency, and coverage or scope of the information and/or its source. In addition to these, some user based criteria are also discussed including reputation, endorsement, consistency, self-confirmation, expectancy violation and persuasive intent heuristic. The study has a special focus on the use of cognitive heuristics in credibility judgement where the authors studied the role of user perception and judgement on credibility results. The suggested factors in the above mentioned categories are listed in the table below

***Table 2.10*. Credibility categories and factors used by [26]**

Sr. No	Category	Factors
1	Accuracy	Error free website, verifiable information
2	Authority	Author, author's credentials, qualifications, whether the site is recommended by a trusted other.
3	Objectivity	The author's purpose of information, if information is fact/opinion, information has commercial intent/conflict of interest
4	Currency	Last update date of information
5	Coverage	Information completeness and depth
User based (subjective) categories and factors		
6	The reputation heuristic	Familiar sources are believed to be more credible than unrecognized sources independent of information quality or source credentials by users.
7	The endorsement heuristic	Sites recommended by known others, or recommended in testimonials, reviews, or ratings are trusted more without content verification
8	The consistency heuristic	Information consistency among different sites. This is superficial validation as agreement with one other independent person or source fulfils it
9	The self-confirmation heuristic	Information is considered credible if it validates users' beliefs even if its well-argued and researched
10	The expectancy violation heuristic	Failure to meet user expectation by a website deems it to be non-credible. Asking or providing more information than necessary or requested, Poor aesthetics and professionalism results in strong negative credibility evaluations
11	Persuasive intent heuristic	Biased information is automatically judged as non-credible by users. E.g. unexpected advertisement

To accurately judge credibility, a range of activities are required to do by users from considering visual design elements and structure of a website to more strenuous information verification. However, users often choose to engage in these superficial ways of aesthetics dependability with least required effort than an effort to verify content or source because users tend to not spend a long time on a given site hence they develop quick judgement strategies.

Thus, the experimental studies in this research argue that aesthetics (section 2.1.2.3) alone does not play an important role in credibility finding, but forms only a part of it.

This study [42] designed a prototype of the tool which works on four credibility judgement criteria i.e. type of website, date of update, sentiment analysis and a pre-defined Google page rank. Each link document in the initial query result set is assigned a weight based on different criteria using the Potentially All Pairwise RanKings of all possible Alternatives (PAPRIKA) method. A separate module to check total number of 'internal links', 'external links' and 'broken links' defined as link integrity is developed. To evaluate the tool performance, its scores were compared with scores given by human judges' which resulted in a low correlation (0.484). This might be because of the tool's restriction to four criteria only and the difference in opinion between human judges about the various influencing factors.

To overcome this limitation, the authors added two more significant web credibility factors in the tool WebCAST, i.e., reputation and review based on users' rating reflecting personal experience with the website. They suggested in [17] that an ideal tool should incorporate all 17 factors indicated in their research review. However, it is difficult to compute all quantitatively. Hence, they use major six factors. The empirical evaluation of this updated tool resulted in a correlation between tool-generated and human judges' score of 0.89. The positive and higher correlation verifies the validity of the tool

In the proposed system in [43] five major areas of website trustworthiness are discussed i.e. Authority, Related resources, Popularity, Age and Recommendation. Eighteen factors are categorized under these five categories. The objective of the proposed system is to provide more trustworthy websites as top results which would save considerable amount of searching time.

Table 2.11 **Credibility categories and factors used by [43]**

Sr. No.	Category	Factors
1	Authority	Page title, meta keyword, meta description
2	Age	last modified date and domain age
3	Popularity	Google PageRank, alexa rank
4	Related Links	Google, yahoo, bing, alexa inbound links
5	Recommendation	Alexa rank, WOT rating, siteadvisor rating, dmoz listing, Google, Yahoo and Bing indexed pages

Authors in [44] claim that majority of existing solutions for web credibility are done by computers or users. However, the user based judgements are costly, time consuming or need expert training. As for computer based judgements, the system lacks a unified model. Therefore, they present a hybrid model combining many factors for evaluation done by computers or humans. They presented all possible factors based on their research mapped into suitable categories and use different evaluation techniques prior to judgement. This hybrid method is very effective in producing reliable results by considering multiple categories.

For organizing the content in given space, the categories and factors are presented in tabular form below.

Table 2.12**. Credibility categories and factors used by [44]**

Sr. No	Category	Factors
1	Accuracy	Source of content, author details, references cited to scientific data, peer review supported by evidence, social and heuristics approach (if majority agrees with the answer or is endorsed by an expert on the topic), use of digital watermarks shows authenticity
2	Authority	Author’s qualification and credential in the web community, author’s contact details, number of articles’ citation, prior contributions and awards received.
3	Aesthetics	Combination of colors, layout, images, videos, fonts, use of bulleted lists, or presentation of tabular data which is consistent on all pages of the website.
4	Professionalism	Presence of advertisements, privacy policy, mission statement or objectives, data protection certification mechanisms, seals, and marks. If it has spelling

		errors, broken links, no multi-language support, Domain name or URL suffix, credentials of members on editorial board, the process taken for maintaining quality of content and often follow the “paid access to information” policy,
5	Popularity	The website traffic or web user’s past experience with the website, social factors like good and bad reviews of the website, ranking in search engine output, ratings given by qualified authors
6	Currency	Presence of up to date information, date stamp/time indicator frequency of update for content
7	Impartiality	Checking biasness of content by collecting positive and negative responses of the given query, Content being peer-reviewed by a group of experts
8	Quality	Includes the reviewer’s experience, ranking of the journal/proceeding

The study in [34] employs machine learning techniques: different regression algorithms like Logistic Regression, Linear Regression, and Support Vector Regression for the model that works on nine web credibility evaluation factors. Two different labelling such as binary labelling (1, 0) and numeric labelling (five point scale rating) are used for defining credibility. The performance is evaluated using 10-fold cross-validation based on Correlation Coefficient (CC), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE). The Logistic Regression, Linear Regression and Support Vector Regression reaches the co-efficient correlation of 0.896, 0.875 and 0.816 respectively. Thus, the Logistic Regression model outperforms other prediction algorithms. The factors are shown in the table below

***Table 2.13*. Credibility categories and factors used by [34]**

Sr. No	Category	Factors
1	Readability	Content visibility and presentation, reading speed, legibility and effort for user.
2	Authority	Control the publisher by comparing it with other websites rankings.
3	Accessibility	Working together of components like web technologies, web browsers, authoring tools, and other user agents, and websites to remove barriers preventing interaction with or easy access to websites.

4	Understandability	Avoiding overly complex sentences, terminology and providing clear layout and design.
5	Popularity	Information-rich and attractive webpage will have high link popularity. The more ratings a post gets, the more reliably the ratings tell the value.
6	Freshness	Up-to-date/latest content especially for tending topics like sports, awards, breaking news etc.
7	Broken Link	Presence of a broken link or dead (link on a web page that no longer works) preventing outside access
8	PageRank	PageRank – Google's algorithmic method to rank pages
9	Duplicate content	Repetition of content at more than one page detected by its location on the page.

Websites are made for human use, therefore, the authors in [45] suggest that behaviour of humans shall be a qualifying factor to evaluate website credibility. The approach considered several factors that could be quantified and collected those by web analytics tools to model the human behaviour in web credibility. The four factors are Average Time of user on website, Pages/Visit (single session), Average Daily Visits of website, Bounce Rate (Number of users who have left the website within ten seconds of their arrival). They used Custom Search API, traffic API, engagement API, and rank and reach API to capture the factors values. Used LDA (Linear Discriminant Analysis). This same experiment was run using WOT and WebCast tools which achieved a positive correlation of 0.87 and 0.71 respectively.

The authors in [46] argue against the use of aesthetics for website credibility judgement by users. The descriptive criteria such as visual presentation can be masked by web templates making the websites look more reliable without underlying content verification. Instead more robust normative criteria can make correct assessments. Thus, as mentioned earlier in this section that aesthetics alone cannot contribute to credibility. This study proposed four objective credibility criteria namely, authority, currency, accuracy and relevance and disregard any subjective factors from other researches. The suggested factors in these categories are shown in the table below

***Table 2.14*. Credibility categories and factors used by [46]**

Sr. No	Category	Factors
1	Authority	Author's qualification, experience, contact details, affiliation , reputation and recognition (author or organization name), position, title (e.g. Dr or Professor), brief detail of the content creator's experience, organization's physical address, web address URL
2	Accuracy	Grammatically correct, no typos (peer reviewed), editorial process, reliable links of editorial process e.g. has the con-tent passed peer-review or has it been reviewed by others
3	Currency	Content's publishing date and last modification date
4	Relevance	Title, type of information, literature, number of citations, frequency of its reference in other documents, publication medium (e.g. book, journal, article, blog, etc.), Content overview (e.g. title, abstract, etc.), References list

The quantitative analysis results suggested that 10 elements were deemed useful for helping and 3 elements were not useful to evaluate the trustworthiness of information. Therefore, the former 10 factors were included and latter 3 were rejected in the proposed credibility criteria.

The authors in [33] propose a framework for website credibility judgement based on customer perspective with a focus on customers' with no awareness in the embedded technologies, business practices and legal grounds behind on-line purchasing. The categories and their factors are mentioned in the table below

***Table 2.15*. Credibility categories and factors used by [33]**

Sr No	Category	Factors
1	Web-site General Appearance (colour scheme, graphic details, embedded service objects and novelty of design ideas)	Design and colour scheme – the overall design, graphical and text elements, colour combinations Unification of elements – the unity of design elements for all pages), Identity integrity – the clear identification of company's activities Brand creative works – company's logos, slogans, corporate colours

		Space usage – the placement of all the site elements in relation to the browser window
2	Personal Information on Executives	CEO's and responsible officers representation, contact information and portraits
3	Company's General Information Accessibility	Company's history
4	Company's Financial Information Accessibility	Standard Annual report on the last year Annual reports on previous five years Quarterly reports on the last year, including the latest finished period report
5	Business Partners and Affiliates Representation	Company's partners and affiliates presentation Links to the company's partners and affiliates extensive information
6	Newsletter	Presence of newsletter
7	Site Navigation Convenience	Clarity of navigation Site tree presentation Major features placement Domain identity, the company's possession of the domain name of certain level. The higher the level of domain, the better the comprehension of address information and easiness of search. Domain name of second level (www.company.***) is the most common case.
8	Logistics integrity	On-line and mail delivery possibility
9	Connection and Loading Speed	Avoidance of large graphic images and embedded applications to avoid long loading time
10	Customer Support Integrity	Utilization of embedded mailing system and web forms

11	Customer Feedback and Complain Response	Response within 1,4 or 24 hours
12	Business to Customer Personalization Rate	Presence of the customer's name in the appeal, the company's officer's name in the signature, the company's contact information, a direct telephone number, a return e-mail address, not a machine-generated reply
13	Multiple Browsers Compatibility	Site appearance in major browsers Applications compatibility to major browsers
14	Contact Information Ease of Use	Contact information and email placement Mailing forms usage for customer convenience
15	Support of Languages	Equal support of all used languages
16	Information on Secured Protocols Utilization	Declaration of secured protocols utilization
17	Trusted Third Parties Seal Presence	Presence of a TTP token from one of the major operators
18	Legal Grounds	Presence of disclaimer, basic terms and conditions, cancellation terms, Reference to customer's protection authority

Most of these factors are good for transparency which builds trust between the customers and the organization. Also, approval of security tokens like TTP by corporate regulations is approved for companies only after an extensive analysis of the company's activities. Thus, its presence increases customer's trust dramatically providing a sense of security. These factors are targeted for commercial websites only, so we eliminate the ones not applicable.

Another research [47] proposed a credibility assessment algorithm that uses seven categories with multiple factors where all the categories are considered equal. The average score of each of the category is the web credibility score. This algorithm was tested on top of an existing QA system where answers are ranked by their credibility score. The research conducted extensive quantitative tests on 211 factoid questions, taken from TREC QA data from 1999-2001. These are

***Table 2.16*. Credibility categories and factors used by [47]**

Sr. No	Category	Factor
1	Correctness	TF-IDF score, Google search rank
2	Authority	(presence of author name, contact information
3	Currency	last update date
4	Professionalism	domain type, Alexa median load time percentage, Google speed score, Mozscape domain authority, Mozscape page authority, WoT trustworthiness users' rating, WoT child safety users' ratings, and WoT experts score
5	Popularity	Web page's share count from multiple social media websites, popularity rank, and traffic rank of the web page
6	Impartiality	sentiment score
7	Quality	readability of the content and its originality to rate its quality

The findings of this study show significant improvement in answer accuracy by four credibility categories including correctness, professionalism, impartiality, and especially quality. Quality and impartiality stood out the most and improved PerCorrect percentage to 5-6% while authority, currency and popularity did not significantly perform better than baselines. The authors claim that their research is the first to propose a web based QA system module to cover all seven categories along with ranking answers based on the category scores.

2.4. Research gap

This literature review shows that most of the state-of-the-art methods require uniform initialization of source trustworthiness and sources and claims infer the truth iteratively from one another. This implies a linear relation between sources and their claims. This impacts the precision, accuracy and convergence rate of the methods affecting their performance.

With assigning uniform weights among all sources, the performance of many truth discovery algorithms relies on the majority types of sources. This strategy can work well with majority good sources but this is not the case in reality.

The research investigation further shows the strengths and limitations of the current techniques. They perform well in general but are unstable most of the time. Most of the researches use only one or a few of the credibility categories in finding the truth, hence, this study shows that there is no one-fits-all solution and a single method did not consistently work better than others. To estimate value veracity, all of the possible categories must be incorporated.

The subjective part of credibility i.e. user’s perception plays a role in credibility but is rarely used. Many user based factors depend on the experience of users where they form their own criteria of trusting information. For e.g. credibility is influenced by information accuracy and surface features categories when users have more experience or expertise in relevant domains, and have high information skills respectively. Thus, by incorporating user judgement, trust scores can be improved substantially [45].

There are many researches supporting or are against use of user based factors. Therefore, we will practically evaluate whether and how much it actually contributes in credibility. Many factors come under user based categories. However, due to the limitation of scope of this research, factors that are infeasible or unable to quantify with direct measures will not be included in this research.

2.5. Research gap tables

Table below shows research gap of truth finding studies using ***single*** credibility categories

***Table 2.17*. Research gap table for single credibility categories**

Sr. No	Research	AC	AU	AE	PR	PO	C	Q
1	Majority Voting [24]							
2	Average-Log [25]							
3	Investment [25]							
4	J. Han et al. [6]	✓						
5	L. Berti-Equille et al. [27]	✓						
6	J. M. Kleinberg [14]		✓					
7	X. Lin et al. [28]		✓					
8	F. Lius et al. [29]		✓					
9	S. M. Shariff et al. [22]			✓				
10	D. Robins et al. [32]			✓				
11	J. Han et al. [7]					✓		
12	X. S. Fang et al [13]					✓		
13	X. S. Fang et al [35]					✓		
14	X. L. Dong et al [36]						✓	
15	X. L. Dong et al [37]						✓	
16	A. Galland et al [11]							✓
17	M. Wu et al [38]							✓

AC – Accuracy AU – Authority AE – Aesthetics

PR – Professionalism PO – Popularity C - Currency

Q – Quality

The table below shows research gap of truth finding studies using ***multiple*** credibility categories

***Table 2.18*. Research gap table for multiple credibility categories**

Sr. No	Research	AC	AU	UB/ AE	PR	PO	C	Q	I
1	M. J. Metzger et al. [26]	✓	✓	✓		✓	✓		✓
2	S. Aggarwal et al. [42]		✓			✓	✓		✓
3	S. Aggarwal et al [17]		✓	✓		✓	✓		✓
4	S. Ansari et al. [43]		✓			✓	✓		
5	A. Ali Shah et al. [44]	✓	✓	✓	✓	✓	✓	✓	✓
6	R. Manjula et al. [34]		✓	✓	✓	✓	✓	✓	
7	H. Singal et al. [45]			✓					
8	J. Pattanaphanchai et al. [46]	✓	✓		✓	✓	✓	✓	
9	V. A. Tsygankov [33]		✓	✓	✓				
10	A. A. Shah et al. [47]	✓	✓		✓	✓	✓	✓	✓

AC – Accuracy AU – Authority UB – User Based
AE – Aesthetics PR – Professionalism PO – Popularity
C – Currency Q – Quality I - Impartiality

CHAPTER 3

Proposed Methodology

3. Proposed Methodology

The main focus of this work is to build an efficient and accurate predictive model to evaluate the web content credibility for resolving conflicts based on leading factors. In the previous section, various factors were indicated that need to be considered for credibility assessment. An automated tool that considers all these factors would be ideal; however it is infeasible given the difficulty of quantitatively measuring all these factors and often there are no direct measures to quantify them. Hence, this study has automated credibility scoring using a few major factors that can be assessed quantitatively as well as automatically. However, these factors cover all the mentioned seven content credibility categories. In addition, the aesthetics category will be evaluated to identify its role in web credibility. As opposed to the previous methods, the source-claim relational dependency cannot be represented by linear functions and is often unknown apriori. Hence, this complex dependency relationship will be learnt by a machine learning classifier such as neural network or decision trees without any prior knowledge on this relationship. Neural networks and decision trees have been used widely in different domains to estimate complex functions with large number of inputs. Hence, these are suitable for this computation. The credibility categories used in this study are defined the categories in table 3.1. The factors used under each category along with their APIs used are discussed in section 3.1.3.

***Table 3.1*. Credibility categories used in this study**

Category name	Explanation
Accuracy	Correctness of website content
Authority	Experience and popularity of the website
Aesthetics	Visual appearance of websites
Professionalism	Efficiency of a website with its policies and features available on the website
Popularity	Website’s reputation among web users and reviewers
Currency	Frequency of update of website content
Quality	Rating of the information in terms of its readability

This section is further divided into two sub-sections; system design and evaluation settings. Different phases of the proposed system from data collection to building classifier will be discussed in the system design sub-section. Various measures used to evaluate the system will be discussed in the evaluation settings sub-section.

3.1. System Design

The bare bones of the system design section involve different phases such as data collection and pre-processing, feature extraction, and building the machine learning classifier. The architecture of the proposed methodology is shown in Fig 3.1.

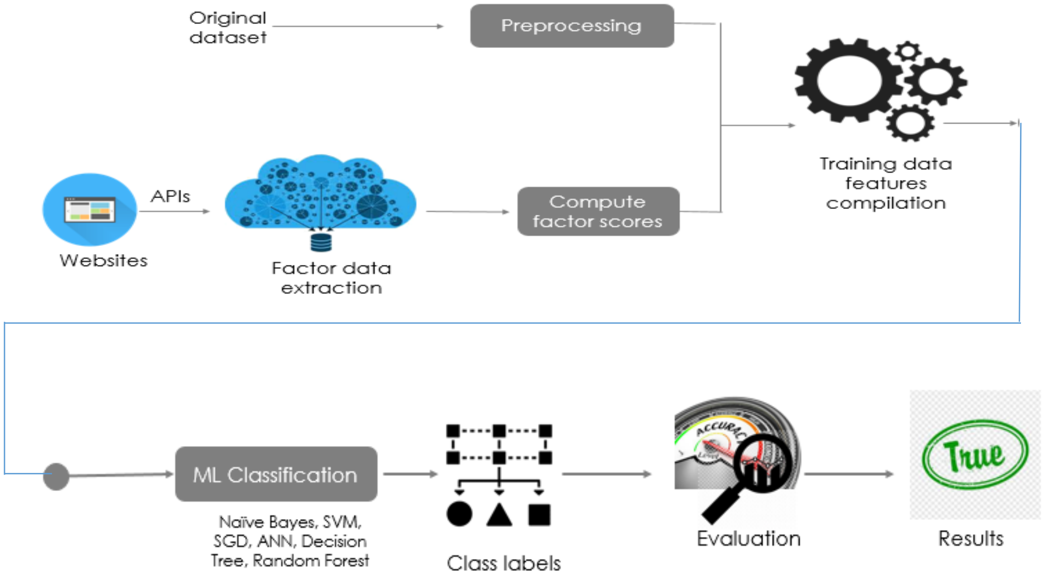


Figure 3.1 **Architecture of the proposed methodology**

3.1.1. Data collection and extraction

This research worked on the Book-Author dataset. It was obtained from the authors of [28]. In its original form, the dataset categorized books into 18 different genres. Each genre file contained 21 columns including ISBN, title of the book, its authors name listed by the seller, seller name, their website and other details etc. This dataset includes the ground truth for a portion of listed books. The dataset was processed to meet the requirements of this research. Firstly, all genre files were combined into one to apply important pre-processing steps. Scattered data was clustered based on identical ISBN number.

After identifying the conflicting authors for different books and using the ground truth list from those, the final number of rows in the dataset was 4061. This contained around 137 books reported by 712 sellers. On average, each book was listed by 4-9 sellers. Among these, the instances reporting the correct author names for a book according to the ground truth are 2691 (class label: yes) and incorrect author names according to the ground truth are 1370 (class label: no).

3.1.2. Text pre-processing

An essential step in Natural Language Pre-Processing (NLP) is text pre-processing. It removes inconsistency, ineffectiveness and noise from the data to transform it in a more understandable and intelligent structural form. Three pre-processing steps were performed on data; tokenization, special character removal and normalisation.

The dataset contained authors listed by different sellers. To extract conflicting author names, they were processed. Firstly, two phase tokenization was applied. In the first phase, since author's names were listed in a single line text, each author name was separated from another based on ‘and’, ‘space’, ‘;’, ‘&’ and ‘,’. In the second phase, the first, middle and last name of each author was broken down. Some of the names had a prefix or suffix like Mr. or Jr. which were removed. All special characters and nuances like [signed] were also removed. Normalisation was also applied to the authors' names by case folding to lower case letters which was utilized in the conflicting author names identification module. This data was saved in another file containing ISBN, book name, seller reporting it, their website URL and author's names by each seller cleaned up and broken down.

In the next step, a matching algorithm for conflicting author names identification was used. It reads names of authors for the same book reported by different sellers and identifies the conflicting ones. *Cosine similarity* was used to find same names with different spellings used where a threshold of 0.3 or 30% was set. Any names less than the threshold value are considered conflicting. Equivalence class cases such as Lily James and Lily J. were also handled. The results achieved were books with conflicting author names reported by different sellers. The portion of this file for which the ground truth was available was used as final data.

- **Tokenisation**

  Sentences can be split at different locations in the sentence like white spaces or commas etc. to transform it into separate words. This is known as tokenisation. It is essential to tokenise sentences before implementing advanced pre-processing steps on data such as stop words or special character removal and lemmatisation. Two phase tokenisation was applied on the authors' names to separate each author and break it down by its first, middle and last name to be used in the conflicting author identification module. The author name was tokenised at 'white space', 'and', 'comma', 'semicolon', and '&'. URLs of seller websites were not tokenised to preserve the correct page link.

- **Stop words removal**

  Words in English language sentences with no special use and frequent occurrence such as articles, prepositions and conjunctions are less useful in identifying the context. Removing stop words reduces the size of data as well as makes data handling easy. Stop words were not removed specifically from the book names and URLs as they are small and each word is important to identify the correct book name. Removing stop words would lead to training the system on improper names and URLs which need to be fixed.

- **Special Character Removal**

  To clean the data from unusual characters that were not part of the actual names, special characters were removed from the authors' and sellers' names. These were not removed from URLs as they are fixed. The nuances like [signed by] were also removed.

- **Normalisation**
  - **Case folding**

    All the author names were transformed to lower case in order to compare author names with each other to extract the conflicting names respectively.

- **Equivalence classes**

  All the author names were compared to find the conflicting authors reported by the sellers for each book. Equivalence class cases like Lily James and Lily J. where both of these names are exactly the same were also handled using cosine similarity.

After pre-processing steps, data was stored in separate file. Factors for each website were then extracted which were used as features for the classifier.

3.1.3. Feature extraction

The performance of web content credibility is greatly impacted by feature extraction. This study collected around 100 different factors contributing towards credibility from multiple researches in this field. Then, feature reduction technique was applied to find the suitable factors among those for the Book-Author dataset used in this study. For example, factors like use of digital watermarks and sentiment features were excluded as the book seller's websites did not contain any articles to extract such information. This resulted in a set of 35 factors. Furthermore, some of the selected factors collectively contributed to one factor score and thus, were reduced. For example, the traffic data of a website from last 3 months was aggregated to a single score by taking their average. The final no. of factors that the system worked on is 24. Multiple APIs and information processing techniques were used to extract data for the selected factors/features under each category. The details of factors, their APIs and feature reduction in each category are given below.

A. Accuracy

Two factors are used in the Accuracy category i.e. Correct data and data richness. It is important for data to be accurate that it can be verified, free of errors and correct semantically and syntactically in some cases like authors.

The first factor, correct data, is found by taking similarity measure of the original author names listed by the seller for a book with the available ground truth. The original data contains many nuances, special characters, spelling errors and often is not in correct order as per the book hardcopy or ground truth. It is directly obtained from the seller website. A code snippet in Java is used to measure similarity between the two texts considering these errors. If the similarity score is more than threshold for most books by a particular seller, the correct data score is more for that seller. This is based on the truth finding principle by [6] that states 'if a website provides many true values or facts, it is considered to be trustworthy'. If the original text is free of errors, the correct text score is set to 60.

The second factor is data richness which tells the expertise of the seller in a particular genre. All the books listed by the seller are combined based on their genre included in the dataset. The percentage of each genre from all the genres for a seller is computed and the maximum available books from a particular genre is taken as the seller's expertise. This indicates that the seller has more books from a particular genre and hence has more expertise or is data rich in that particular genre. This increases the seller's probability of data correctness in that genre [28]. The value of data richness is computed via Java code. If the correct text instance from factor 1 also belongs to the expertise genre of the seller, the remaining score of 40 is set to the accuracy score.

B. Authority

Two factors i.e. Domain Authority and Page Authority are used in this category. Domain Authority is a score to predict the likelihood of a domain to rank on search engines [48]. It measures this by comparing it with other websites. Like Domain Authority, the Page Authority is a score to predict the ranking strength but of a single page [48][49]. Both of these are not Google ranking factors. The scores range from 1 to 100 where the higher the score, the greater the ability to rank. The values of these factors are derived from Mozscape API.

C. Aesthetics

The factors contributing to aesthetics category are Visual score, expected users to like the site, Visual appearance and Visual clarity. The latter three are contributing to the Visual score so these were excluded as part of feature reduction. The value for Visual score is extracted using the Visual Mind AI engine.

D. Professionalism

Factors used in this category are Domain type, Broken links, Page load time, Page title, meta tags. A sub category of security is included in this category as it caters all those factors and professional websites must contain some sort of security. The security factors include WOT trustworthiness user rating, WOT expert score, WOT child safety rating, spam score and presence of standard security protocol (http/https).

Domain type is extracted from the bookseller’s website URLs. A broken link is a link that no longer works because of an improper URL or a non-existent external webpage to which it is linked. It affects a website’s usability by reducing traffic and damaging rankings as it prevents search engine crawlers from indexing its pages [50]. Presence of a broken link/links has a negative impact on credibility. Its value is derived from ‘nibbler tool’.

Page load time was computed using Java code where each URL was loaded from different IPs at different times and their average was taken. Page title and meta tags including meta keywords and meta descriptions provide key information about a webpage. A professional website must contain titles on all of its pages and meta tags to provide information helpful in ranking the pages. These scores are retrieved from the nibbler tool. Feature reduction is applied here to reduce the factors to most relevant. Page title and meta tags are average to a meta tags score.

The security factors; WOT trustworthiness user rating, WOT expert score and WOT child safety rating are obtained from MyWOT API. It is a Web of Trust – a community experience based platform where millions of users rate websites for their security and trustworthiness. It provides website safety checks based on ratings and reviews from the community and a combination of machine learning (ML) algorithms.

Spam score indicates the likelihood of the website to be penalized or banned by Google based on similar features sites that are penalized or banned. Its value is fetched from Mozscape API. The last factor in this category is the presence of standard security protocol (http/https). Certain websites could not be opened due to security issues as they were lacking the certificates. These websites cannot be checked for their content let alone considered trustworthy or reliable. Its value is obtained using Java code by analysing the URLs.

E. Popularity

The factors used in this category are average 3 months traffic, Global traffic rank, Pages/Visit, Average visit duration, Google PageRank, Inbound links (link popularity).

A website's traffic for the last 3 months is averaged to 'average 3 months traffic'. Additionally, Pages/Visit and Average visit duration contribute to the Global traffic rank. Hence, these were reduced from the features list. The values for these are derived from SimilarWeb API. The next factor; GooglePageRank is obtained via Java code. Inbound links or link popularity is obtained using Mozscape API.

F. Currency

Two factors are used in this category; last update date and Domain age. A website with latest or most recent information and a website published old enough are considered more reliable.

The last update date is retrieved from the nibbler tool. Domain Age defines how old a particular website is. A website lasting long enough reflects its importance and is considered more reliable. Its value is obtained from the WhoIS directory API.

The currency score is assigned as per the range used by [47]. The range is between 0 to 100. Webpages that are created or updated less than a year are given a score of 100. While webpages older than five years are given zero score. Pages with no specified date are also given zero score. All other webpages are given scores between 1-99 as per their dates.

G. **Quality**

Factors contributing to the Quality category are Readability score, Understand-ability score, Accessibility score and Bounce Rate.

Readability allows the system to determine if the content can be easily read and comprehended by most of the users including adults and children. Flesch Kincaid Reading Ease and Flesch Kincaid grade level tests are used to determine this score in the system. These are two widely used measures of Readability. Their value is obtained from Readable API. The average of these two tests provides the Readability score.

Understand-ability score is used to determine that the content is simpler, avoids overly complex sentences, terminology and provides clear layout and design. The Automated Readability Index (ARI) is a measure used to determine the understand-ability of the text. It is obtained by Readable API. For the clear layout and design, the experience value is obtained from the nibbler tool. An average of these two factors provides the Understand-ability score.

Accessibility score is an average of multiple browsers compatibility/mobile friendliness and Page Speed Insights by Google. It determines how accessible the website is to most users including the mobile and disabled users and website's performance on these devices. The browser/mobile compatibility score is obtained from the nibbler tool and Page Speed Insights is obtained by the Page Speed Insights API by Google. An average of these two scores provides the accessibility score.

Bounce Rate is the percentage of users who have left the website within seconds of their arrival or view only one page before leaving. This reflects the quality of a website. A user with better experience and other factors would find the website more reliable and relevant and would usually stay. The value of bounce rate is obtained from SimilarWeb API. The aggregate score of Quality category is the sum of these 4 scores.

The *impartiality category* is not utilized is in study as it does not apply on the type of dataset used. After the application of feature reduction technique, the system is trained on 24 features in 7 categories which are to the point, relevant and suitable to the dataset. A summary of the factors used in each category in this study is shown in the table 3.2

***Table 3.2*. Credibility categories with their factors used in this study**

Sr. No	Category Name	Factors
1	Accuracy	Correct data, Data richness
2	Authority	Domain Authority, Page Authority
3	Aesthetics	Visual Score
4	Professionalism	Domain type, broken links, Page load time, Meta tags score, WOT trustworthiness ratings, WOT expert score, WOT child safety rating, spam score, http/https (presence of standard security protocol)
5	Popularity	Global traffic rank, average 3 months traffic, Google PageRank, Inbound links (Link popularity)
6	Currency	Last update date, Domain age
7	Quality	Readability score, understand-ability score, accessibility score, bounce rate

3.1.3.1. Data transformation - Nominal to numerical entries

Once all the factor values were obtained, any nominal entry was transformed to numerical data. For example, the domain type was transformed as shown in the table 3.3. The domain age was transformed to no. of years. For simplicity, an entry like 10 years 4 months was truncated to 10 years but an entry like 10 years 8 months was rounded off to 11 years. Similarly, the last updated date is transformed to number of days. Then both of these were given a score of bracket range of 0-100 based on their dates.

***Table 3.3*. Domain type - nominal to numeric transformation**

Domain type	Numeric equivalent
Com	1
Dk	2
Org	3
Net	4
co.uk	5
De	6
Ie	7
Biz	8
Scot	9
Edu	10
Ca	11
info	12
com.au	13
co.nz	14
In	15
gov	16
store	17
Site	18

3.1.4. Machine Learning Classifier

We trained our machine on six different classifiers for numerical data namely; Naive Bayes, Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Decision Tree, Random Forest and Neural Network. We did three types of experiments to train the machine. First, all categories were used concretely together. Secondly, each category was added separately to see its contribution in web credibility and thirdly, a combination of categories was tested. More on this can be found in chapter 4. For a dataset of 4061 rows, we used stratified 10 fold cross validation to avoid over-fitting and under-fitting problems.

3.2. Evaluation settings

In order to evaluate the proposed methodology, the metrics are discussed in detail below.

3.2.1. Evaluation metrics

This section discusses the metrics used to evaluate the performance of the framework. Since it is a binary classification problem (i.e. only two classes: yes and no), the metrics used to assess the classifier performance are as follows:

- **Confusion Matrix**

  It is a matrix used commonly to present true classes on Y-axis in the form of true Positive (TP) and True Negative (TN) and the predicted classes on X-axis in the form of False Positive (FP) and False Negative (FN) predictions of the model.

	ACTUAL	
PREDICTED	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

- **Precision**

  It is the ratio of true positives to the sum of true positives and false positives. It tells no of observations correctly classified. Its formula is given in eq. (1).

$$\text{Precision} = TP/\ (TP + FP) \tag{1}$$

- **Recall**

  It is the ratio of true positives to the sum of true positives and false negatives. Its formula is given in eq. (2)

$$\text{Recall} = TP/\ (TP + FN) \tag{2}$$

- **F1 Score**

It is weighted harmonic mean of precision and recall. We fix the value of beta to 1 to favor both precision and recall. Its formula is given in eq. (3)

$$F_{beta} = (1+\beta^2)\frac{precision * recall}{\beta^2 * precision + recall} \tag{3}$$

- **Accuracy**

Accuracy is the fraction of correctly classified prediction with all predictions. It is used to measure how accurately model a performed. Its formula is given in eq. (4)

$$ACC = \frac{tp + tn}{tp + fp + tn + fn} \tag{4}$$

- **Kappa metric**

It measure how much better is your model over the random classifier that predicts based on class frequencies. Its formula is given in eq. (5)

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{5}$$

Where po is observed agreement and pe is expected agreement.

- **ROC Curve**

All the true positive rate and false positive rates are plotted on a chart to visualize trade-off between them using the ROC curve. Classifiers with more top-left side curves are better

CHAPTER 4

Results and Discussion

4. Results and Discussion

In this chapter, the performance of the algorithms will be evaluated. The Book-Author dataset with conflicting author names against different books provided by different sellers was used to evaluate the system performance. A series of experiments were performed using different sets of credibility categories with different features generated for each experiment. In addition to this, two different data split configurations were used to train and test the system. In both configurations, stratified 10 fold cross validation and hold out validation was used respectively to avoid over-fitting and under-fitting problems.

4.1. Classifier performance

The system was trained on six different classifiers for numerical data namely; Naive Bayes, Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Decision Tree, Random Forest and Neural Network. The classifiers were trained using the fetched features for each experimentation. The system performance will be evaluated by comparing the metrics discussed in section 3.2. for all these six classifiers and the one with best performance is selected.

Firstly, all categories were used together concretely in one run. Secondly, each category was individually used in the model to see how accurately the system performs. Thirdly, a combination of categories was tested where each category was added to the previous in each iteration and the results were evaluated.

The dataset consists of 4061 rows with binary classes. It consists of 2691 rows with Y class and 1370 rows with N class. To avoid any over-fitting or under-fitting, stratified 10 fold cross validation and holdout validation was used. In the first type of validation, the whole dataset is divided in 10 equal parts where 9 parts are used for training and the 10th part is used for testing in each iteration. This is repeated 10 times with each part being tested once.

Another set of training and testing was performed with hold out validation where a portion of data was removed from the dataset to be used as an unseen test set. The remaining data was used to train and validate the system using 10 fold cross validation. Then, the unseen test set was applied to see the system performance.

Each of these experiments from the first data split were performed on all of the six classifiers. The classifier with the best performance or better accuracy was chosen. Random Forest proved to be performing well in all the situation and hence, it was selected for our methodology. Then, the second data split configuration (hold out validation) is used once on all categories to test the system performance and shows how accurately it delivers the results.

4.1.1. Experiment 1: All categories

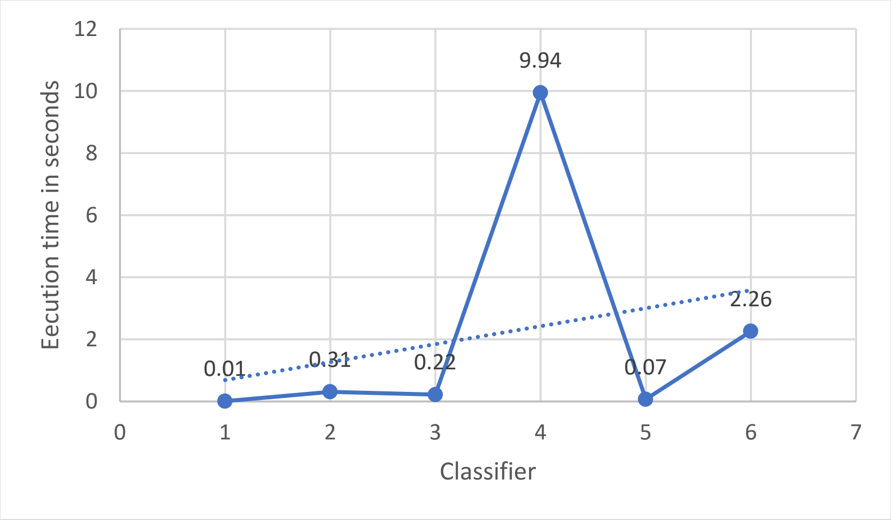
In the first experiment, all the seven categories are used concretely to train the classifiers. The evaluation of each classifier is discussed below.

Table 4.1. **Evaluation metrics of classifiers using all categories**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic
Naïve Bayes	60.42%	0.656	0.604	0.615	0.2063
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	74.19%	0.761	0.742	0.700	0.3162
Neural Network	94.70%	0.947	0.947	0.947	0.8807
Decision Tree	92.90%	0.929	0.929	0.929	0.8398
Random Forest	**97.46%**	**0.975**	**0.975**	**0.974**	**0.9424**

Table 4.2 **Execution time of algos. Each algo is run with 15 epoch iterations and 100 batch size as required**

Sr. No	Classifier	Executional time (seconds)
1	Naïve Bayes	**0.01**
2	SVM	0.31
3	SGD	0.22
4	Neural Network	9.94
5	Decision Tree	**0.07**
6	Random Forest	2.26



1 = Naïve Bayes	2 = SVM	3 = SGD
4 = Neural Network	5 = Decision Tree	6 = Random Forest

Figure 4.1 **Execution time (in seconds) of classifiers with 15 epoch iterations**

As it can be seen in the table 4.1, the top three performing classifiers are Neural Networks, Decision Tree and Random Forest with Random Forest giving the best accuracy score of 97.46%. The precision, recall and f-measure for Random Forest is on top. It is highlighted in bold. The misclassification ratio of Naïve Bayes is high as its performance is unreliable with many complex features. However, the execution time of Naïve Bayes is the lowest as shown in table 4.2 and figure 4.1. SVM and SGD still perform average with an increased execution time making them unsuitable for this problem. Decision Tree has a remarkable execution time with a great increase in accuracy. Neural Network and Random Forest are the top 2 performing classifiers. However, the difference in their execution time is huge making Neural Network computationally very expensive for an accuracy that is closer to the best. For a problem where better computational complexity with good accuracy is required, Decision Tree would be the best option.

Naïve Bayes has used the least no. of resources and is lightweight giving the best execution time other classifiers but the trade-off between the execution time and accuracy for Naïve Bayes and Random Forest is not comparable. The linear dotted line in figure 4.1. shows the execution time trend of the classifier where Neural Network is an outlier. Based on these metrics, we choose Random Forest as the classifier for our methodology with the best accuracy and good efficiency.

It is also used for the second data split configuration discussed in section 4.1.4. The confusion matrix and ROC curves for the classifiers are given below for further reference of results.

Confusion Matrix

The confusion matrix specifies the number of correctly and incorrectly classified instances. Confusion Matrices for all the six classifiers are given below. The first column is a and second is b where a = N and b = Y class. It can be seen that the best classification is done by the Random Forest.

Table 4.3 Confusion Matrices for experiment 1

888 (TP)	482 (FP)
1125 (FN)	1566 (TN)

Confusion Matrix for Naive Bayes

50 (TP)	1320 (FP)
0 (FN)	2691 (TN)

Confusion Matrix for SVM

411 (TP)	959 (FP)
89 (FN)	2602 (TN)

Confusion Matrix for SGD

1241 (TP)	129 (FP)
86 (FN)	2605 (TN)

Confusion Matrix for Neural Networks

1199 (TP)	171 (FP)
117 (FN)	2574 (TN)

Confusion Matrix for Decision Tree

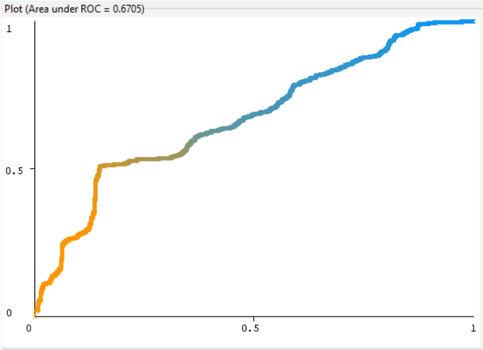
1274 (TP)	96 (FP)
7 (FN)	2684 (TN)

Confusion Matrix for Random Forest

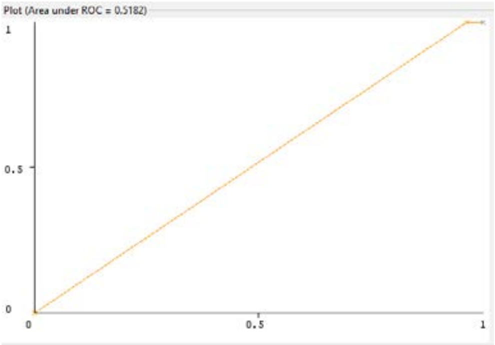
ROC Curve

The ROC plots true positive and false positive rate and evaluates the performance at all classification thresholds. The Roc curve for each classifier is given below. It can be seen that the Random Forest ROC Curve is higher on the upper left corner indicating its good performance.

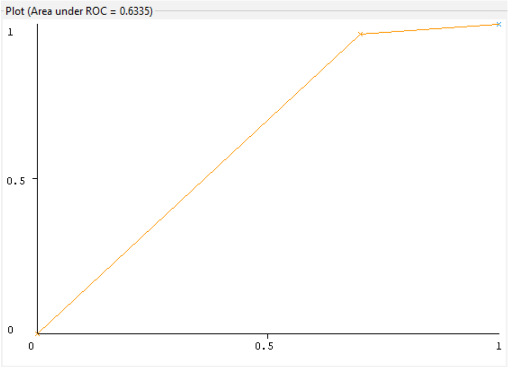
The x-axis shows the false positive rate and y-axis shows the true positive rate



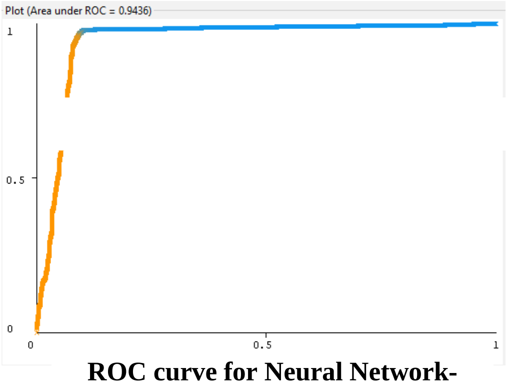
ROC curve for Naive Bayes - experiment 1



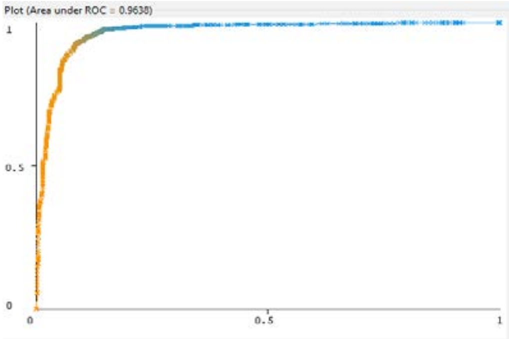
ROC curve for SVM - experiment 1



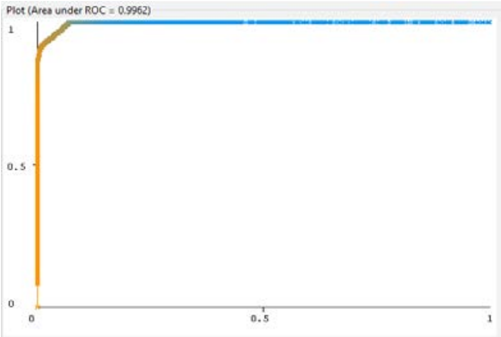
ROC curve for SGD- experiment 1



ROC curve for Neural Network- experiment 1



ROC curve for Decision Tree - experiment 1



ROC curve for Random Forest- experiment 1

Figure 4.2 ROC curves for the classifiers – experiment 1

4.1.2. Experiment 2: Category wise evaluation

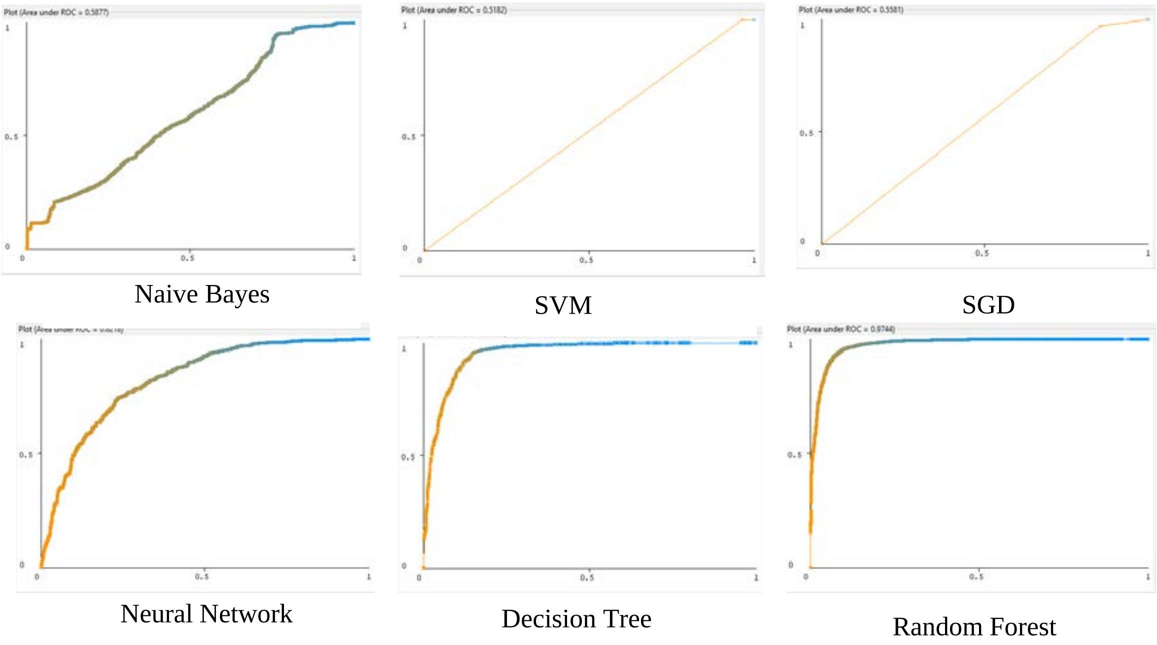
In this experiment, each category is individually used to train the classifier. The evaluation metrics of all the six classifiers used for each category are discussed below.

1. **Accuracy**

   The system is trained using the Accuracy factors. It can be seen from the table 4.4. that Decision Tree and Random Forest performed on top with Random Forest giving the highest accuracy of 93.12%. This is very less in comparison to experiment 1 which uses all categories together. Performance of Neural Network has drastically decreased with only one factor which shows its reliability with many features. The rest of the classifiers perform at an average and almost similar to experiment 1.

Table 4.4 **Evaluation metrics of classifiers using Accuracy category**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic
Naïve Bayes	70.25%	0.712	0.703	0.637	0.1841
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	69.24%	0.700	0.692	0.616	0.1446
Neural Network	77.27%	0.766	0.773	0.766	0.4661
Decision Tree	91.45%	0.914	0.914	0.918	0.8153
Random Forest	**93.12%**	**0.931**	**0.931**	**0.931**	**0.8516**



Naive Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

Figure 4.3 **. ROC curves for the classifiers – experiment 2: Accuracy**

2. Authority

By training the system using factors from the authority category, the following results were achieved.

Table 4.5 **Evaluation metrics of classifiers using Authority category**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic
Naïve Bayes	66.26%	0	0.663	0	0
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	66.26%	0	0.663	0	0
Neural Network	66.19%	0.535	0.662	0.529	-0.0005
Decision Tree	92.85%	0.929	0.929	0.928	0.8366
Random Forest	**95.29%**	**0.953**	**0.953**	**0.953**	**0.893**

Table 4.6 **Execution time of algos for Authority category with only 2 factors**

Sr. No	Classifier	Executional time (seconds)
1	Naïve Bayes	**0.03**
2	SVM	0.44
3	SGD	0.12
4	Neural Network	1.22
5	Decision Tree	0.14
6	Random Forest	2.86

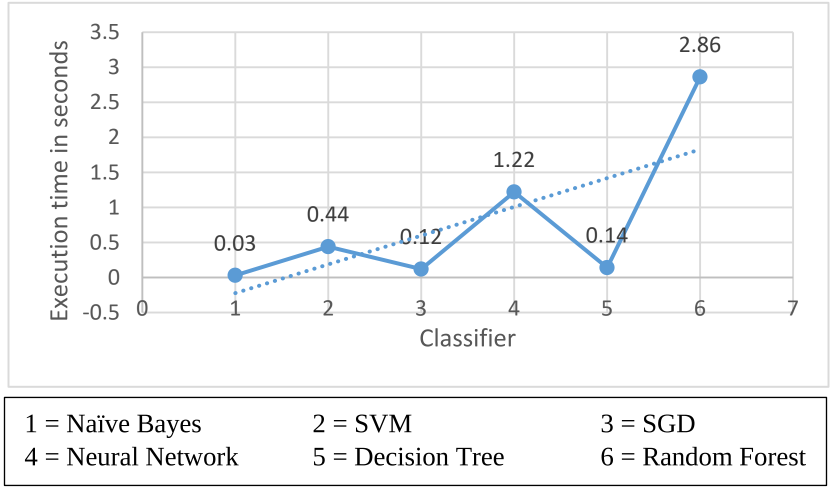
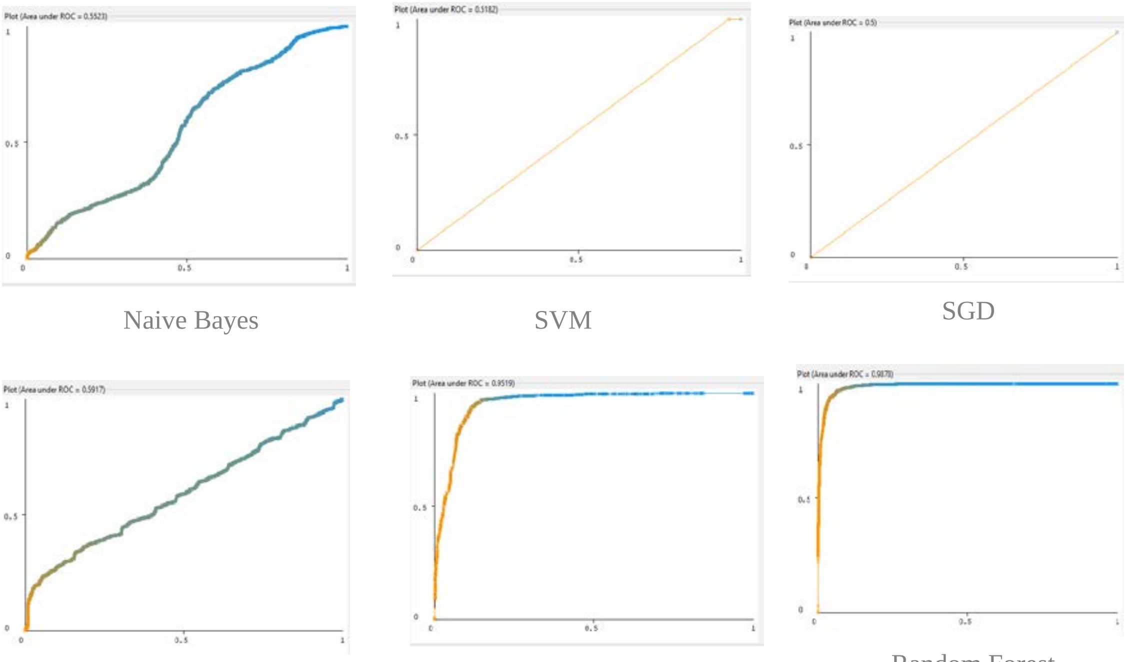


Figure 4.4 **Execution time (in seconds) graph of classifiers - Authority category**

As it can be seen in Table 4.5, Decision Tree and Random Forest perform the best with top accuracy of 92.85% and 95.29% respectively. Naïve Bayes and SVM perform very poorly with this category as their accuracies are average but execution time (Refer to table 4.6.) has increased using only 2 factors as compared to experiment 1 with 24 factors. However, Neural Network's execution time has drastically decreased making it computationally cheap but the accuracy has also decreased a good amount. The execution time for Decision Tree has doubled with no increase in accuracy. As for Random Forest, it again performs the best with highest accuracy while its execution time remains similar to experiment 1. Though the accuracy is still less than collective categories. Thus, it can be seen that using authority category alone decreases the accuracy in comparison to experiment 1.



Naive Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

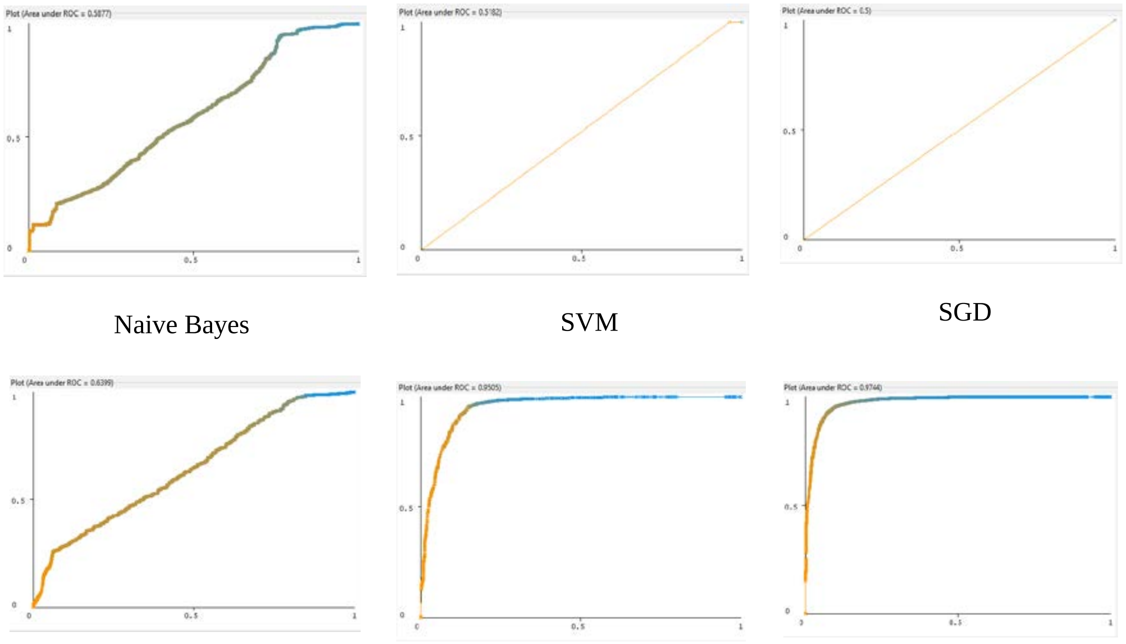
Figure 4.5 **ROC curves for the classifiers – experiment 2: Authority**

3. Aesthetics

By training the system using factors from the aesthetics category, the following results were achieved. As it can be seen in the table below that Random Forest performs the best. However, the accuracy and hence, precision, recall and F-measure have further decreased as compared to experiment 1 and using aesthetics category alone. This answers Research Question 4 that aesthetics play an important role in website credibility, but, its contribution might be less than other categories. When used in combination with other categories (like experiment 1), it gives best results.

Table 4.7 **Evaluation metrics of classifiers using Aesthetics category**

Classifiers	**Accuracy**	**Precision**	**Recall**	**F measure**	**Kappa statistic**
Naïve Bayes	70.25%	0.712	0.703	0.637	0.1841
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	66.26%	0	0.663	0	0
Neural Network	70.47%	0.724	0.705	0.636	0.1854
Decision Tree	91.89%	0.919	0.919	0.918	0.8153
Random Forest	**93.47%**	**0.935**	**0.935**	**0.934**	**0.8516**



Naive Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

Figure 4.6 **ROC curves for the classifiers – experiment 2: Aesthetics**

4. Professionalism

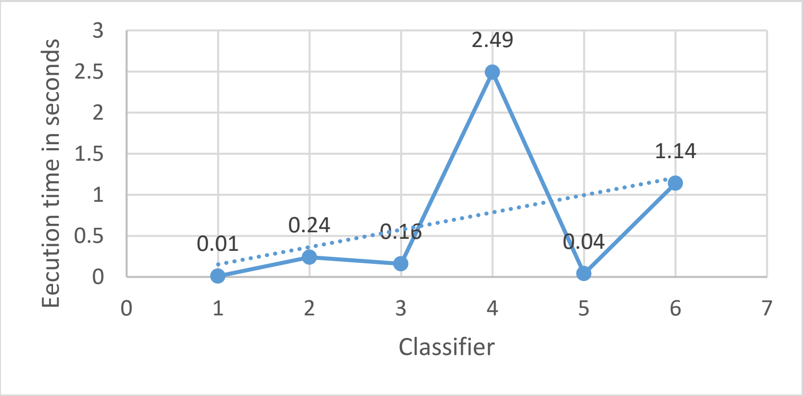
By training the system using factors from the professionalism category, the following results were achieved.

Table 4.8 **Evaluation metrics of classifiers using Professionalism category**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic
Naïve Bayes	61.80%	0.602	0.618	0.608	0.1089
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	66.26%	0	0.663	0	0
Neural Network	78.89%	0.795	0.789	0.791	0.5385
Decision Tree	93.22%	0.932	0.932	0.932	0.8472
Random Forest	**97.02%**	**0.971**	**0.970**	**0.970**	**0.9325**

Table 4.9 **Execution time of algos for Professionalism category with 9 factors (maximum)**

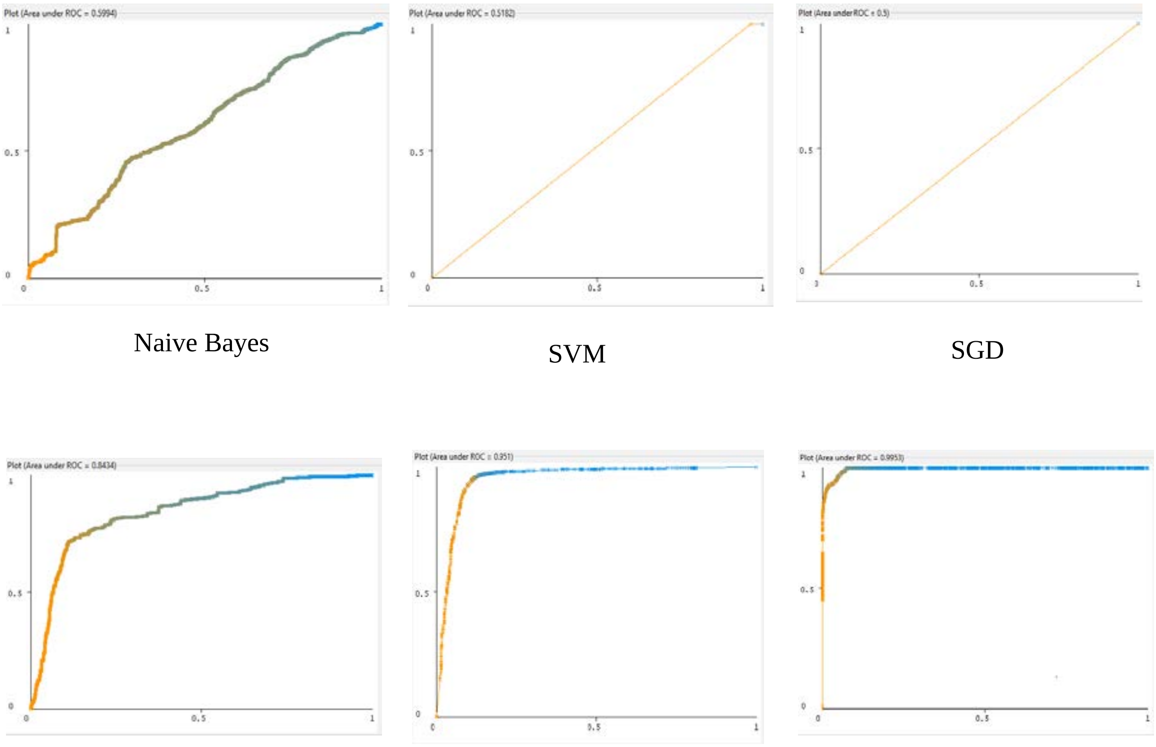
Sr. No	Classifier	Executional time (seconds)
1	Naïve Bayes	**0.01**
2	SVM	0.24
3	SGD	0.16
4	Neural Network	2.49
5	Decision Tree	0.04
6	Random Forest	1.14



1 = Naïve Bayes	2 = SVM	3 = SGD
4 = Neural Network	5 = Decision Tree	6 = Random Forest

Figure 4.7 **Execution time (in seconds) graph of classifiers - Professionalism category**

As it can be seen from table 4.8 and 4.9, with increasing the no. of factors compared to Authority category, Naïve Bayes again gives the lowest execution time (which is exactly same as experiment 1 with all categories) with an average accuracy (which is almost same as or with a minor increase in accuracy than experiment 1). SVM and SGD show no difference in accuracy as compared to Authority category but execution time decreases for SVM while it increases for SGD than experiment 1. Neural Network accuracy has decreased a lot but at the same time its execution time has lowered a great no. This category alone performs better than the rest with Random Forest accuracy close to experiment 1. Its best performance can be seen in the ROC curve as well. However, SGD and Neural Network performance has greatly decreased than any other category so far.



Naive Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

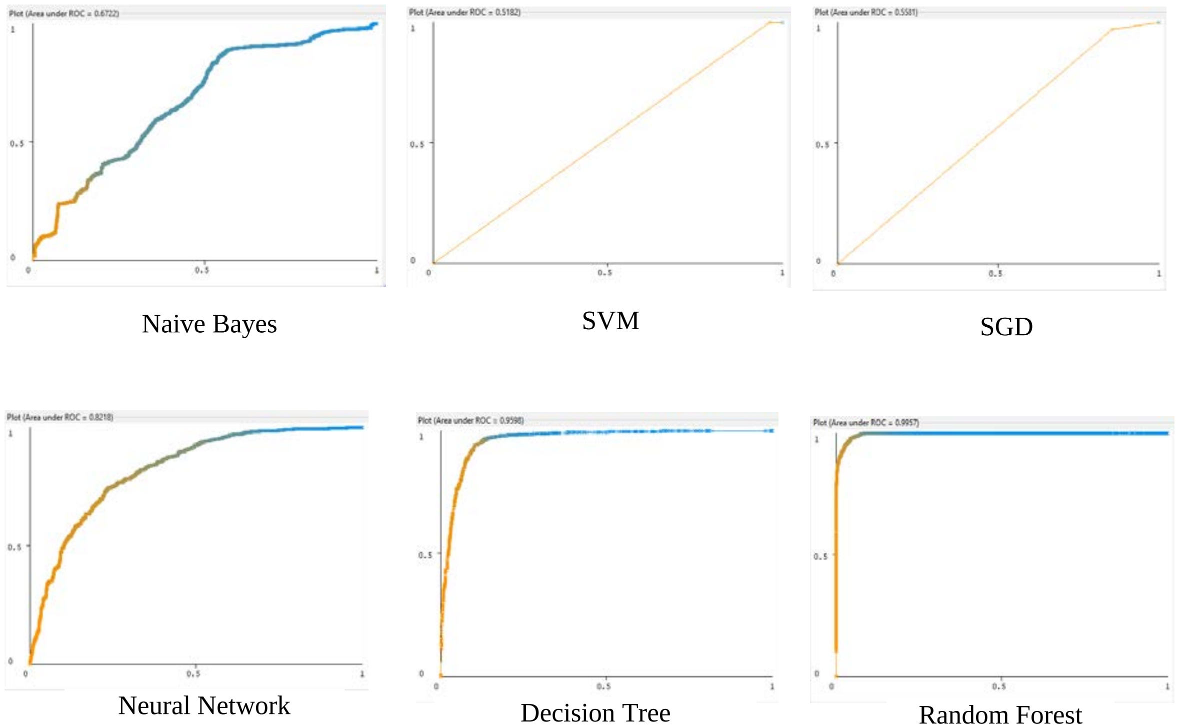
Figure 4.8 **ROC curves for the classifiers – experiment 2: Professionalism**

5. Popularity

By training the system using factors from the popularity category, the following results were achieved. All the classifiers except Naïve Bayes have performed similar to the previous category experiment. Naïve Bayes has further reduced its accuracy showing how unstable the simple classifier is in similar situations (with many factors). Separately, Random Forest performs best in these with an accuracy of 97.04. This is also close to experiment 1 using all categories.

Table 4.10 **Evaluation metrics of classifiers using Popularity category**

Classifiers	**Accuracy**	**Precision**	**Recall**	**F measure**	**Kappa statistic**
Naïve Bayes	51.46%	0.662	0.515	0.508	0.1421
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	69.24%	0.700	0.692	0.616	0.1446
Neural Network	77.27%	0.766	0.773	0.766	0.4661
Decision Tree	93.01%	0.930	0.930	0.930	0.8414
Random Forest	**97.04%**	**0.971**	**0.970**	**0.970**	**0.9329**



Naive Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

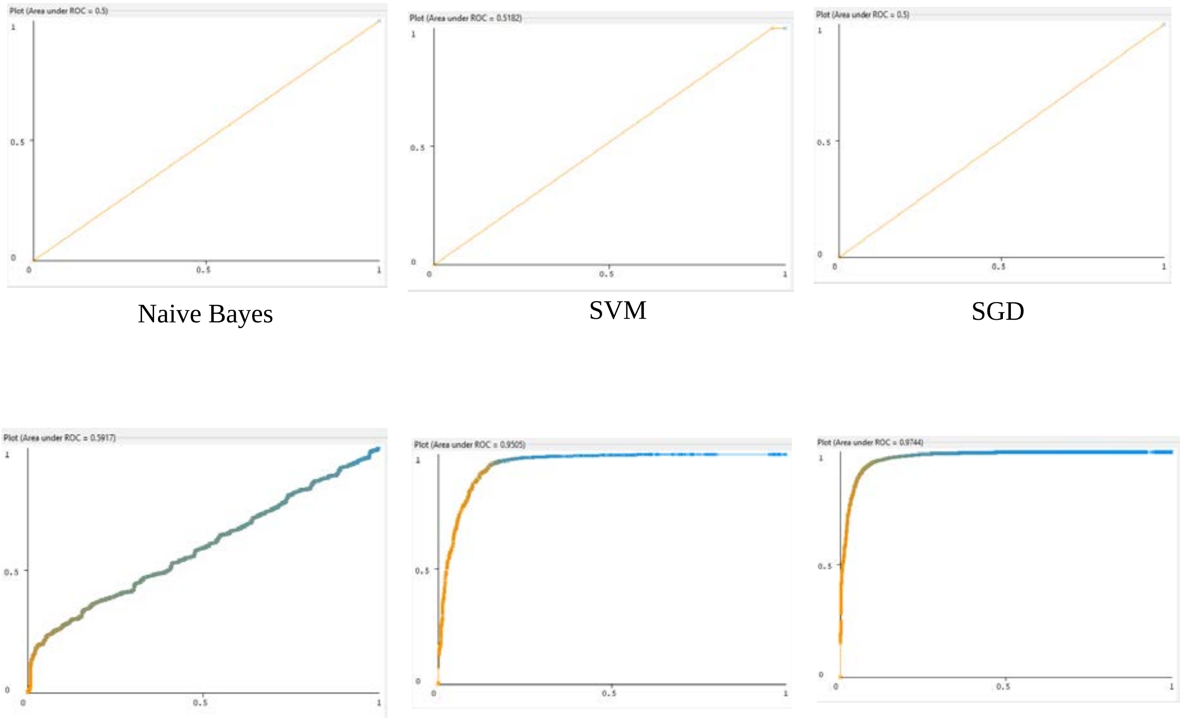
Figure 4.9 **ROC curves for the classifiers – experiment 2: Popularity**

6. Currency

By incorporating Currency factors, following results are achieved. Random Forest performed the best of all these but its accuracy is very less as compared to experiment 1 and also to other categories.

Table 4.11 **Evaluation metrics of classifiers using Quality category**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic
Naïve Bayes	66.26%	0	0.663	0	0
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	66.26%	0	0.663	0	0
Neural Network	67.47%	0.674	0.605	0.679	0.1854
Decision Tree	91.89%	0.919	0.919	0.918	0.8153
Random Forest	**92.47%**	**0.925**	**0.925**	**0.924**	**0.8316**



Naive Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

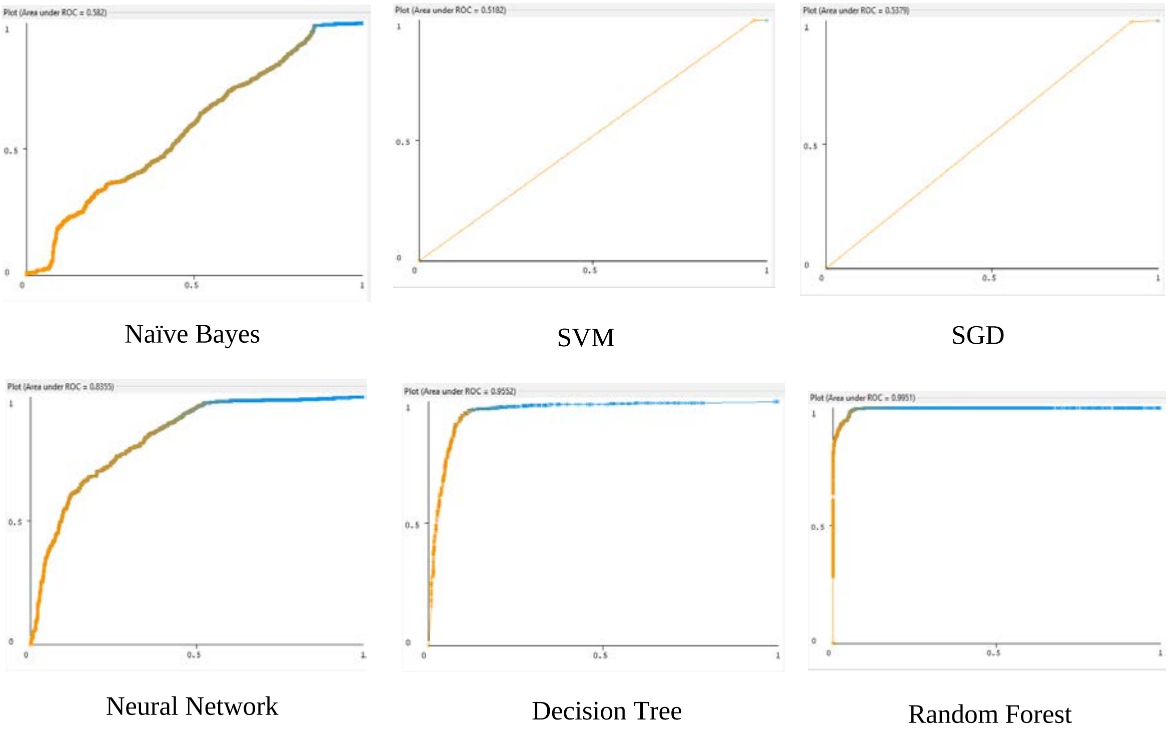
Figure 4.10 **ROC curves for the classifiers – experiment 2: Currency**

7. Quality

By training the system using factors from the Quality category, the following results were achieved. Random Forest performs the best with an accuracy of 97.19% which is also closer but less than experiment 1 with all categories. Naïve Bayes has improved with this category and Neural Network drastically increases in accuracy with these factors than many other categories alone. However, it is still less than experiment 1 with all categories.

Table 4.12 **Evaluation metrics of classifiers using Quality category**

Classifiers	**Accuracy**	**Precision**	**Recall**	**F measure**	**Kappa statistic**
Naïve Bayes	68.77%	0.682	0.688	0.614	0.1368
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	68.72%	0.765	0.687	0.585	0.0978
Neural Network	80.52%	0.821	0.805	0.786	0.5092
Decision Tree	93.37%	0.933	0.934	0.933	0.8503
Random Forest	**97.19%**	**0.972**	**0.972**	**0.972**	**0.9365**



Naïve Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

Figure 4.11 **ROC curves for the classifiers – experiment 2: Quality**

4.1.3. Experiment 3: Combination of categories

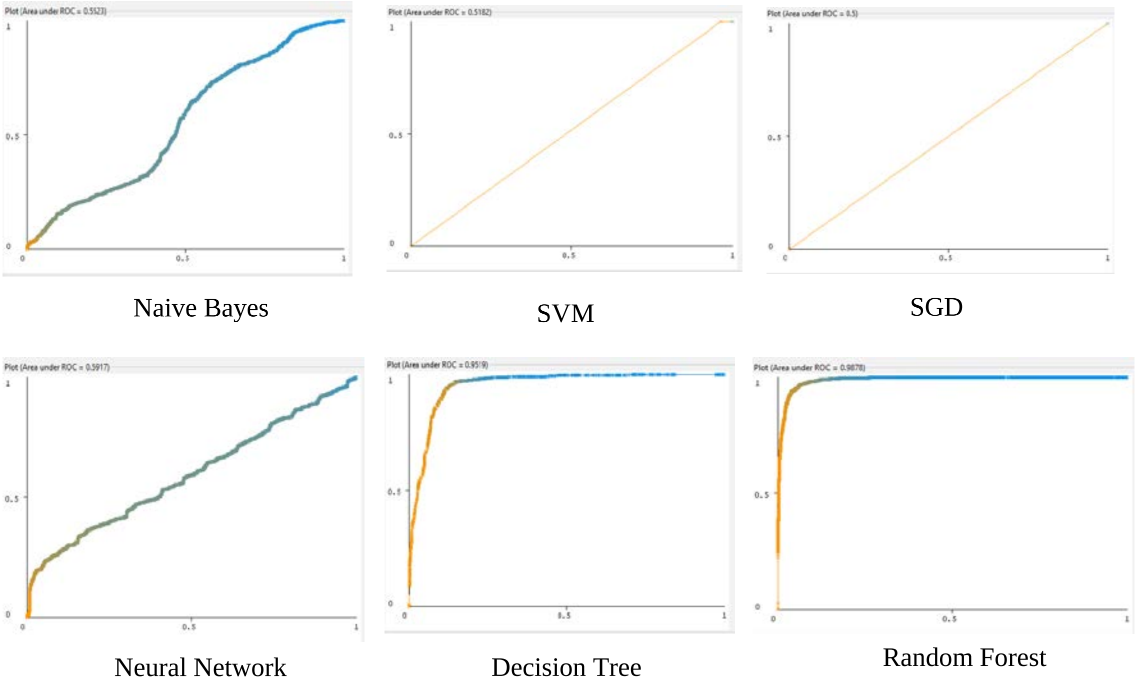
In this experiment, each category is added to the previous ones in each run. The evaluation results from all six classifiers for category combinations are given below

1. **Accuracy and Authority**

   By training the systems using categories from accuracy and authority category, the following results are achieved. All the classifiers perform at an average accuracy except Decision Tree and Random Forest. Random Forest has the highest accuracy which is still less than experiment 1.

Table 4.13 **Evaluation metrics of classifiers using combination 1 categories**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic
Naïve Bayes	66.26%	0	0.663	0	0
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	66.26%	0	0.663	0	0
Neural Network	66.19%	0.535	0.662	0.529	-0.0005
Decision Tree	92.85%	0.929	0.929	0.928	0.8366
Random Forest	**95.29%**	**0.953**	**0.953**	**0.953**	**0.893**



Naive Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

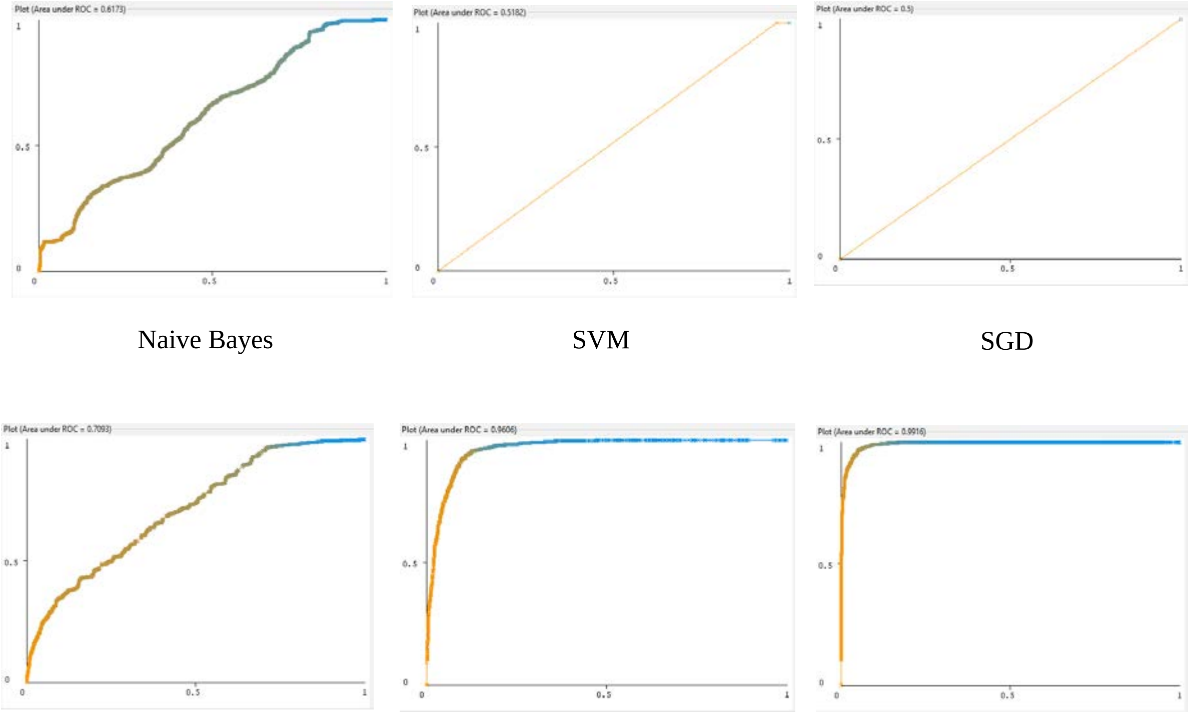
Figure 4.12 **ROC curves for the classifiers – experiment 3: combination 1 categories**

2. **Accuracy, Authority and Aesthetics**

   By incorporating these three categories into the classifier, it can be seen that there is no remarkable difference than previous category combination. However, Random Forest has performed better among these classifiers. In comparison to experiment 1, its accuracy is lower by using only 3 categories.

Table 4.14 **Evaluation metrics of classifiers using combination 2 categories**

Classifiers	**Accuracy**	**Precision**	**Recall**	**F measure**	**Kappa statistic**
Naïve Bayes	69.88%	0.698	0.699	0.637	0.1806
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	66.26%	0	0.663	0	0
Neural Network	73.45%	0.754	0.735	0.689	0.2921
Decision Tree	92.56%	0.925	0.926	0.925	0.8316
Random Forest	**95.91%**	**0.960**	**0.959**	**0.959**	**0.9068**



Naive Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

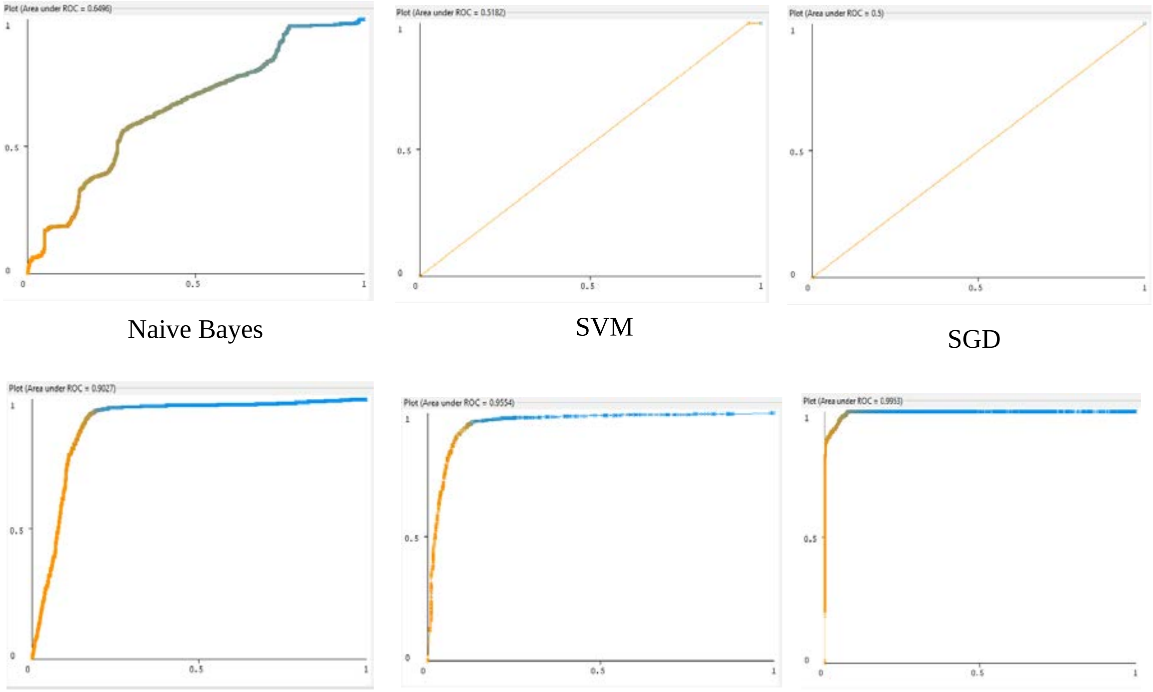
Figure 4.13 **ROC curves for the classifiers – experiment 3: combination 2 categories**

3. Accuracy, Authority, Aesthetics and Professionalism

By using these four categories, there is a sudden increase in Neural Network and Random Forest accuracy while the other 4 classifier perform almost the same as previous categories. Random Forest performs the best with accuracy closer to but less than experiment 1 (using all categories). This shows that Professionalism category highly contributes towards better accuracy for credibility. It could also be seen in experiment 2.

Table 4.15 **Evaluation metrics of classifiers using combination 3 categories**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic
Naïve Bayes	63.82%	0.632	0.638	0.635	0.1763
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	66.26%	0	0.663	0	0
Neural Network	90.49%	0.905	0.905	0.904	0.7824
Decision Tree	92.93%	0.929	0.929	0.929	0.8404
Random Forest	**97.24%**	**0.973**	**0.972**	**0.972**	**0.9374**



Naive Bayes

SVM

SGD

Neural Network

Decision Tree

Random Forest

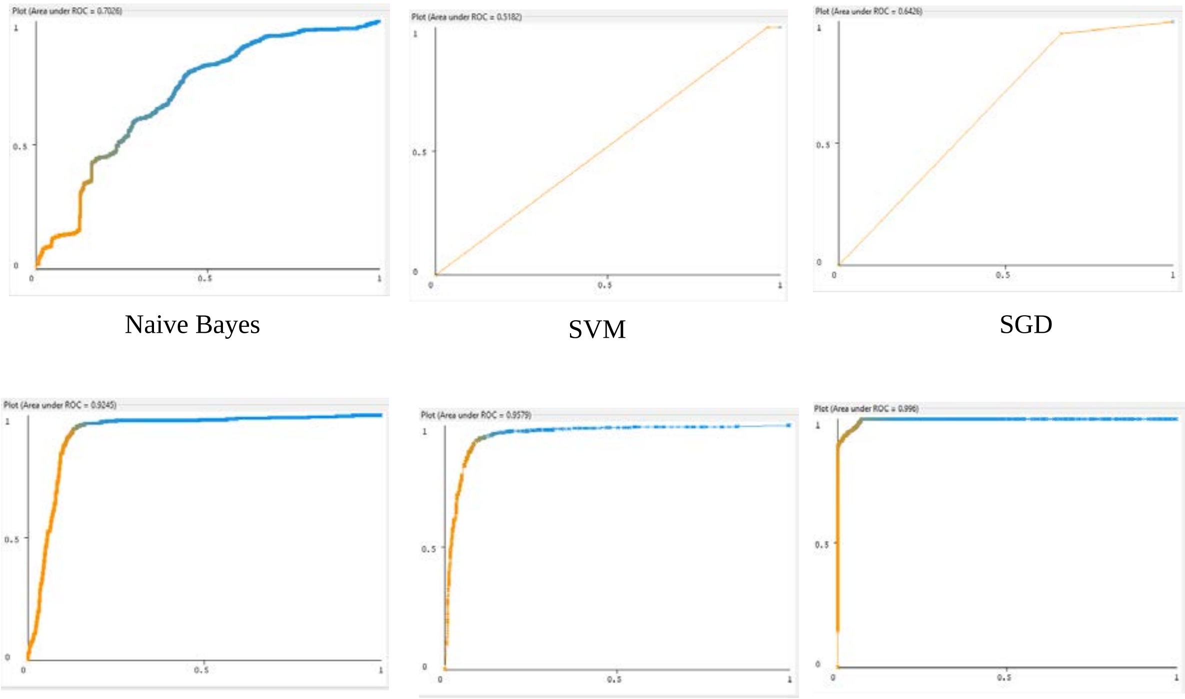
Figure 4.14 **ROC curves for the classifiers – experiment 3: combination 3 categories**

4. Accuracy, Authority, Aesthetics, Professionalism and popularity

When factors from these five categories are used to train the system, the last 3 classifiers perform slightly better than category 3 combinations. However, the performance of Naïve Bayes has declined. Random Forest performs the best with accuracy closer to experiment 1. Like combination 3, this experiment shows popularity also contributes equivalent to professionalism in credibility and better than rest of the categories.

Table 4.16 **Evaluation metrics of classifiers using combination 4 categories**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic
Naïve Bayes	56.73%	0.678	0.567	0.572	0.2006
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	74.26%	0.750	0.743	0.708	0.3314
Neural Network	92.21%	0.922	0.922	0.922	0.8237
Decision Tree	92.83%	0.928	0.928	0.928	0.8397
Random Forest	**97.36%**	**0.974**	**0.974**	**0.973**	**0.9401**



Naive Bayes SVM SGD

Neural Network Decision Tree Random Forest

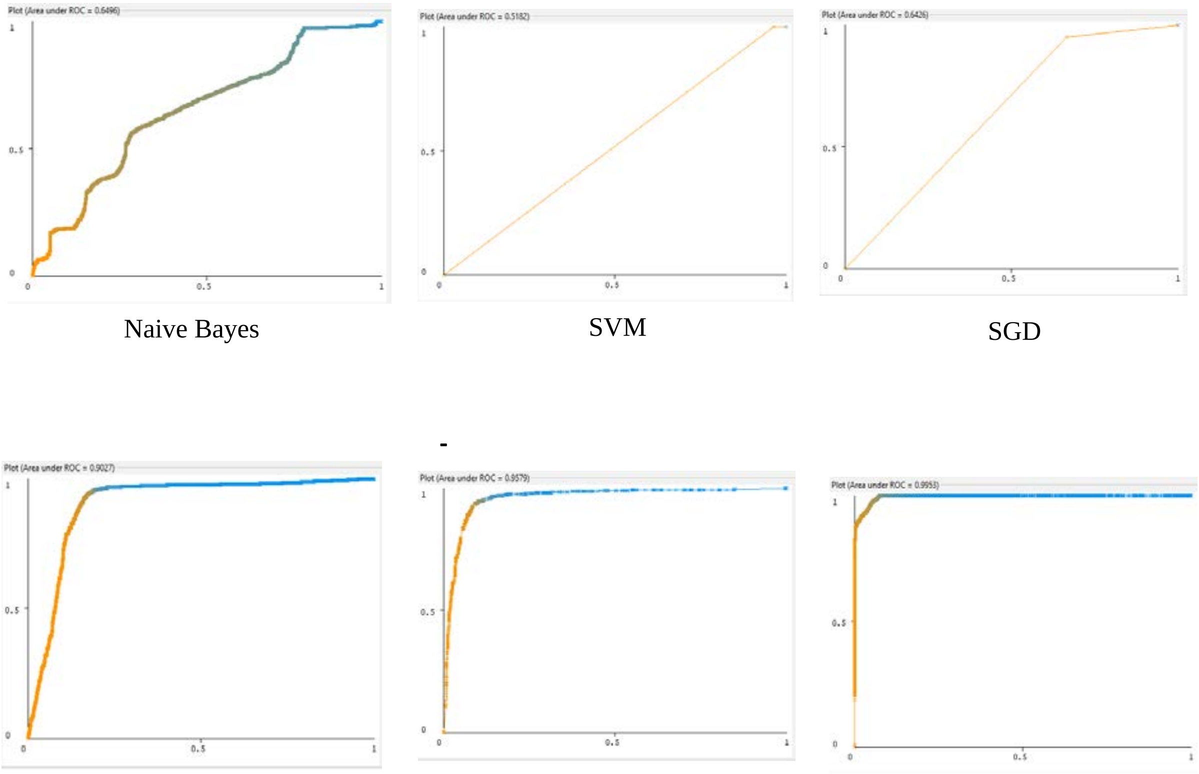
Figure 4.15 ROC curves for the classifiers – experiment 3: combination 4 categories

5. Accuracy, Authority, Aesthetics, Professionalism, popularity and currency

By using factors from these categories, the following results are achieved. It can be seen that Random Forest performs better than rest of the classifiers. All the classifiers provide accuracy close to category 3 combination except Naïve Bayes. But the performance of all is very close to experiment 1 with all categories. Adding the last category is equivalent to experiment 1

Table 4.17 **Evaluation metrics of classifiers using combination 5 categories**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic
Naïve Bayes	63.82%	0.632	0.638	0.635	0.1763
SVM	67.49%	0.782	0.675	0.556	0.0478
SGD	74.26%	0.750	0.743	0.708	0.3314
Neural Network	91.29%	0.912	0.912	0.912	0.7824
Decision Tree	92.73%	0.927	0.927	0.927	0.8404
Random Forest	**97.26%**	**0.972**	**0.972**	**0.972**	**0.9401**



Naive Bayes SVM SGD

Neural Network Decision Tree Random Forest

Figure 4.16 ROC curves for the classifiers – experiment 3: combination 5 categories

4.1.4. Experiment with data split configuration 2 – hold out validation

A second type of data split configuration was used to evaluate the system performance where a portion of data was removed from the training file. The training data was used to train and validate the classifiers. Stratified 10 fold cross validation was used to avoid over-fitting and under-fitting problems. Random Forest is chosen as the best fitted classifier for this problem and hence will be discussed in this section.

Table 4.18 **Evaluation metrics of Random Forest using all categories for data split configuration 2**

Classifiers	Accuracy	Precision	Recall	F measure	Kappa statistic	Execution time
Random Forest	97.64%	0.977	0.976	0.976	0.9463	2.97s

Once the classifier was trained and validated with this date, the unseen test set was supplied. The test set contained 75 instances of a book without the ground truth. The classifier predicted 52 instances with 100% confidence. Upon checking with ground truth, those instances were correctly classified. 3 out of 75 instances were wrongly classified by the classifier with a confidence score of 0.822, 0.822 and 0.724 respectively

Few instances of the author names provided in the data compared with ground truth as per correct classification and misclassification can be seen in the table below

ISBN: 60974176

Book Name: The Machine That Changed the World: The Story of Lean Production

Ground truth: James P. Womack; Daniel T. Jones; Daniel Roos

Table 4.19 **Classified values by Random Forest in data split configuration 2 experiment**

Original text in data	Predicted	Prediction confidence	Actual answer from ground truth
Womack, James P; Jones, Daniel T & Roos, Daniel	2:Y	1	Y
James Womack	1:N	0.983	N
Womack, James	2:Y	0.822	N
James P ; Daniel T	2:Y	0.822	N
James Womack, Daniel Jones, Daniel Roos	1:N	0.724	Y

This experiment proves the accuracy of our classifier based on the given factors and categories.

4.2. Result Discussion

4.2.1. Result discussion for experiments based on accuracy

The results of experiment 1 with all categories is summarized in graphical form in figure 4.16. Firstly, all categories used collectively provide the best results with Random Forest. It gives an Accuracy of 97.46%, Precision 0.975, Recall 0.975 and F-measure 0.974.

When each category is used separately, Professionalism, Popularity and Quality have more contribution than Accuracy, Authority, Aesthetics and Currency.

Furthermore, it can be seen from the results that categories with more number of factors perform well in comparison to the ones with a single or two factors. The Professionalism, Popularity and Quality categories performed very close to experiment 1 done with all categories concretely used. While the rest of the categories, Accuracy, Authority, Aesthetics and Currency did not contribute to credibility as much. These categories consist of a maximum of two factors while the former categories contain many. This shows that a single factor alone such as data richness, domain expertise, domain authority, endorsement, inbound links etc. cannot provide for credible websites accurately. Using these in combination with other factors belonging to the same or different category drastically improves the results. This experiment has improved accuracy than many researches.

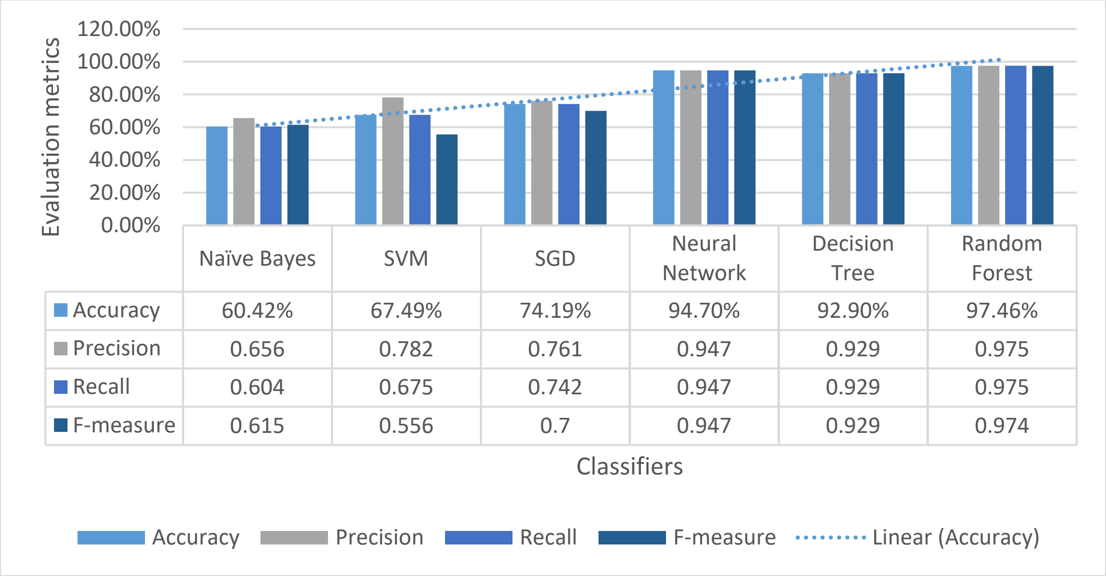
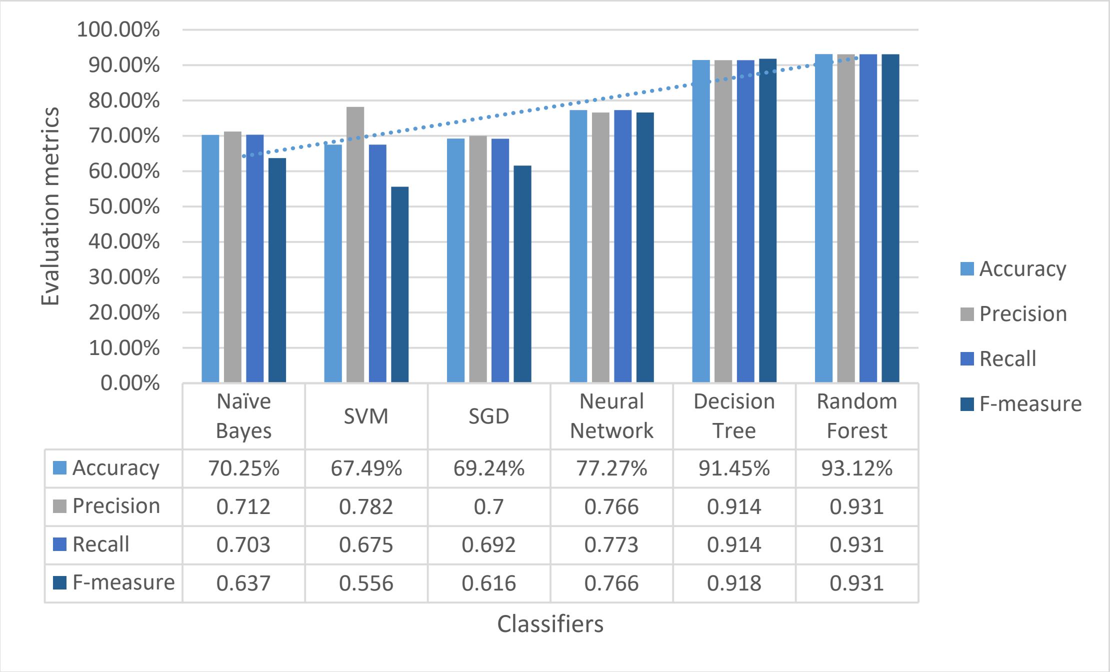
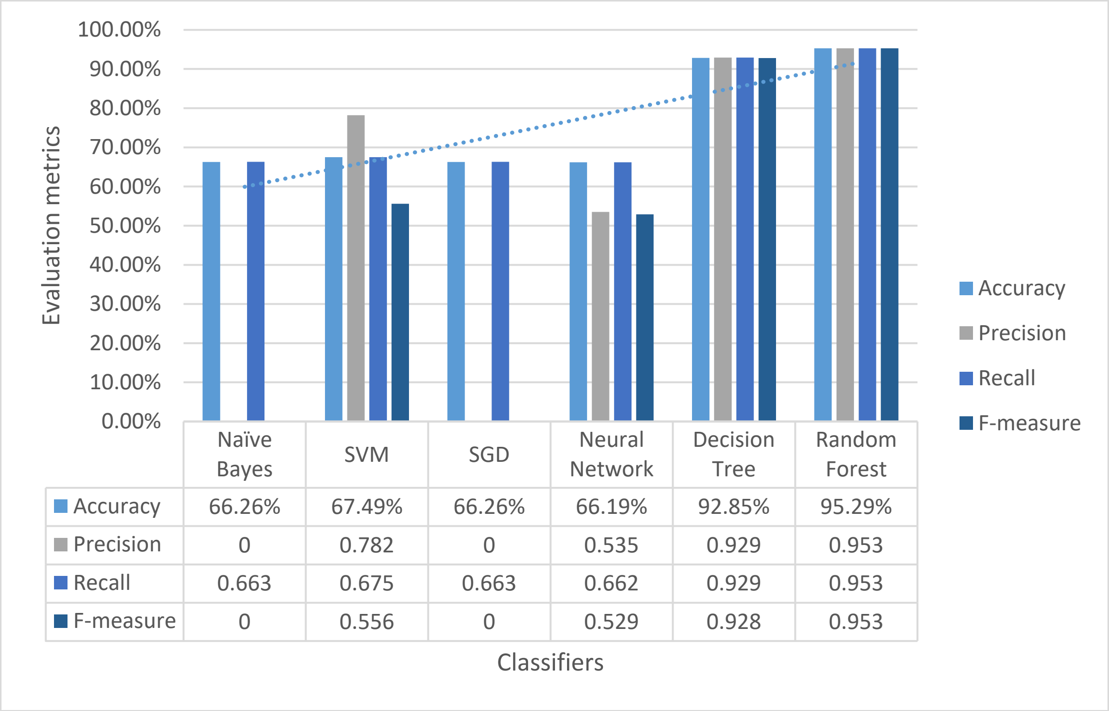


Figure 4.17 **Comparison of evaluation metrics for experiment 1 - all categories**



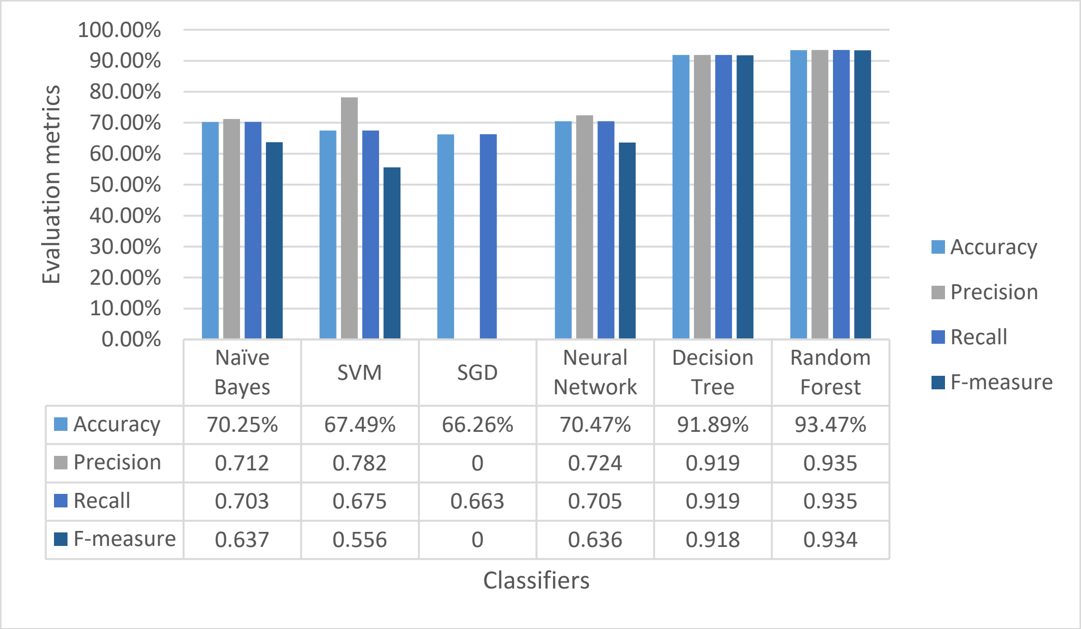
	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	70.25%	67.49%	69.24%	77.27%	91.45%	93.12%
Precision	0.712	0.782	0.7	0.766	0.914	0.931
Recall	0.703	0.675	0.692	0.773	0.914	0.931
F-measure	0.637	0.556	0.616	0.766	0.918	0.931

Figure 4.18 **Comparison of evaluation metrics for experiment 2 – Accuracy**



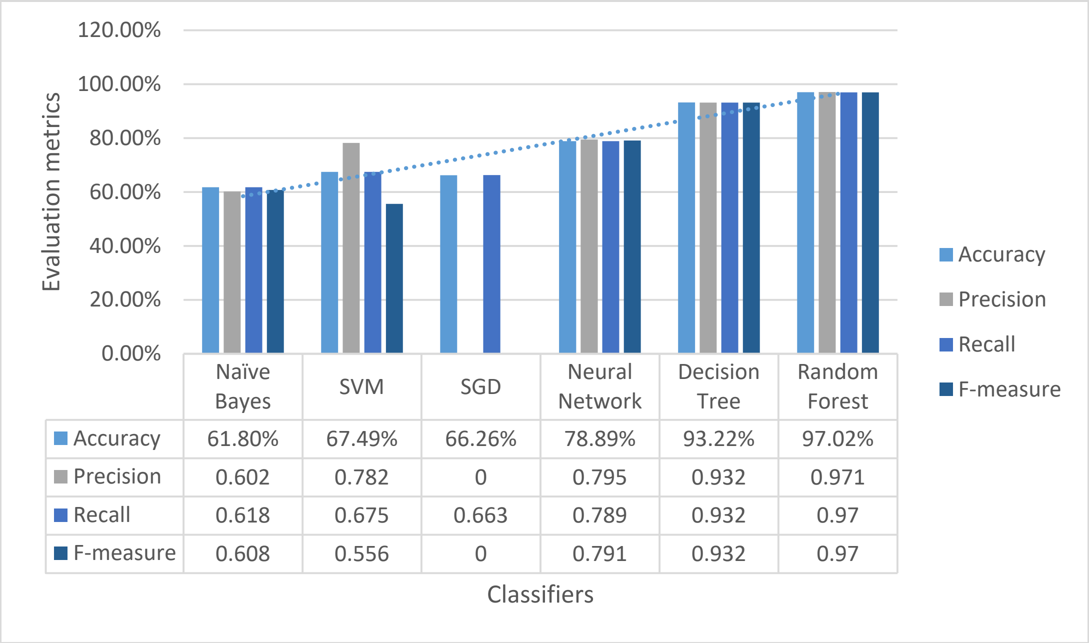
	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	66.26%	67.49%	66.26%	66.19%	92.85%	95.29%
Precision	0	0.782	0	0.535	0.929	0.953
Recall	0.663	0.675	0.663	0.662	0.929	0.953
F-measure	0	0.556	0	0.529	0.928	0.953

Figure 4.19 **Comparison of evaluation metrics for experiment 2 - Authority**



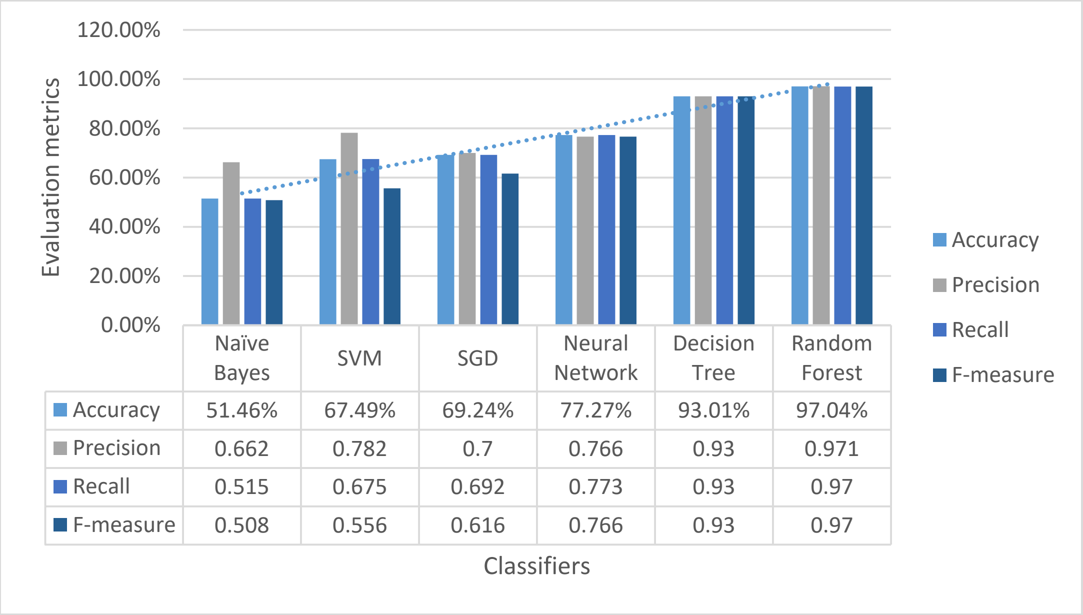
	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
■ Accuracy	70.25%	67.49%	66.26%	70.47%	91.89%	93.47%
■ Precision	0.712	0.782	0	0.724	0.919	0.935
■ Recall	0.703	0.675	0.663	0.705	0.919	0.935
■ F-measure	0.637	0.556	0	0.636	0.918	0.934

Figure 4.21 **Comparison of evaluation metrics for experiment 2 - Aesthetics**



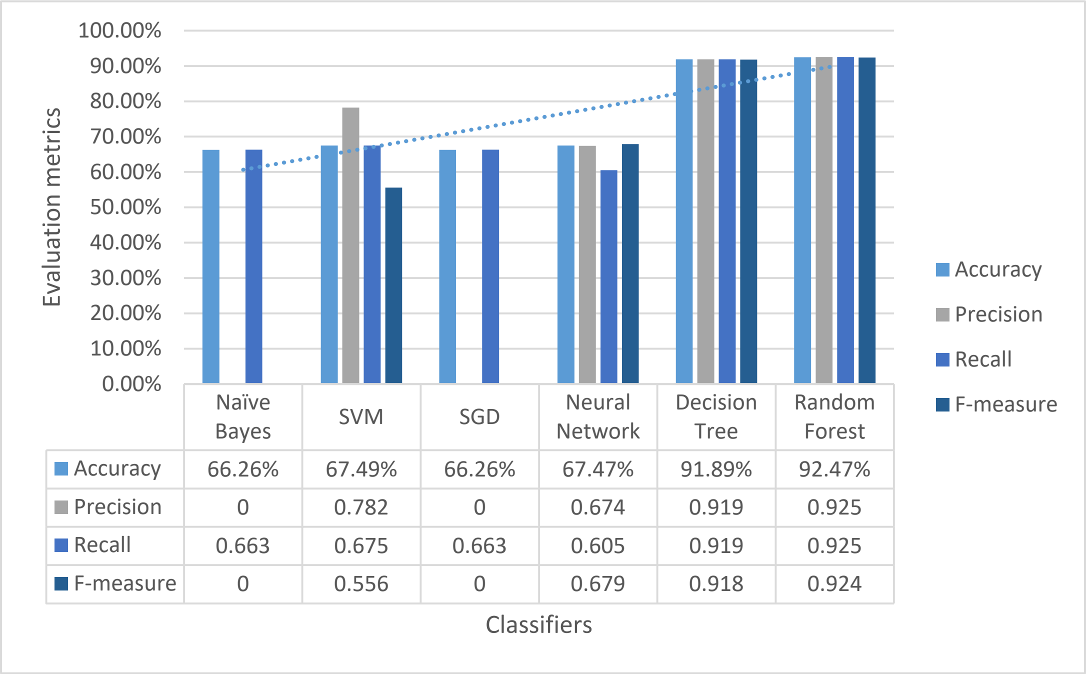
	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
■ Accuracy	61.80%	67.49%	66.26%	78.89%	93.22%	97.02%
■ Precision	0.602	0.782	0	0.795	0.932	0.971
■ Recall	0.618	0.675	0.663	0.789	0.932	0.97
■ F-measure	0.608	0.556	0	0.791	0.932	0.97

Figure 4.20 **Comparison of evaluation metrics for experiment 2 - Professionalism**



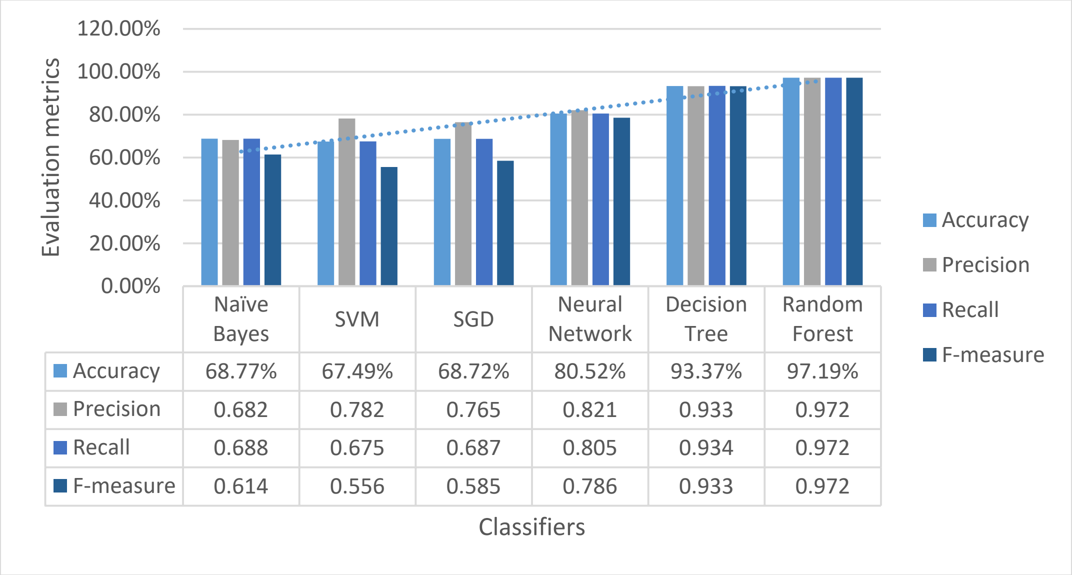
	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	51.46%	67.49%	69.24%	77.27%	93.01%	97.04%
Precision	0.662	0.782	0.7	0.766	0.93	0.971
Recall	0.515	0.675	0.692	0.773	0.93	0.97
F-measure	0.508	0.556	0.616	0.766	0.93	0.97

Figure 4.22 **Comparison of evaluation metrics for experiment 2 - Popularity**



	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	66.26%	67.49%	66.26%	67.47%	91.89%	92.47%
Precision	0	0.782	0	0.674	0.919	0.925
Recall	0.663	0.675	0.663	0.605	0.919	0.925
F-measure	0	0.556	0	0.679	0.918	0.924

Figure 4.23 **Comparison of evaluation metrics for experiment 2 - Currency**



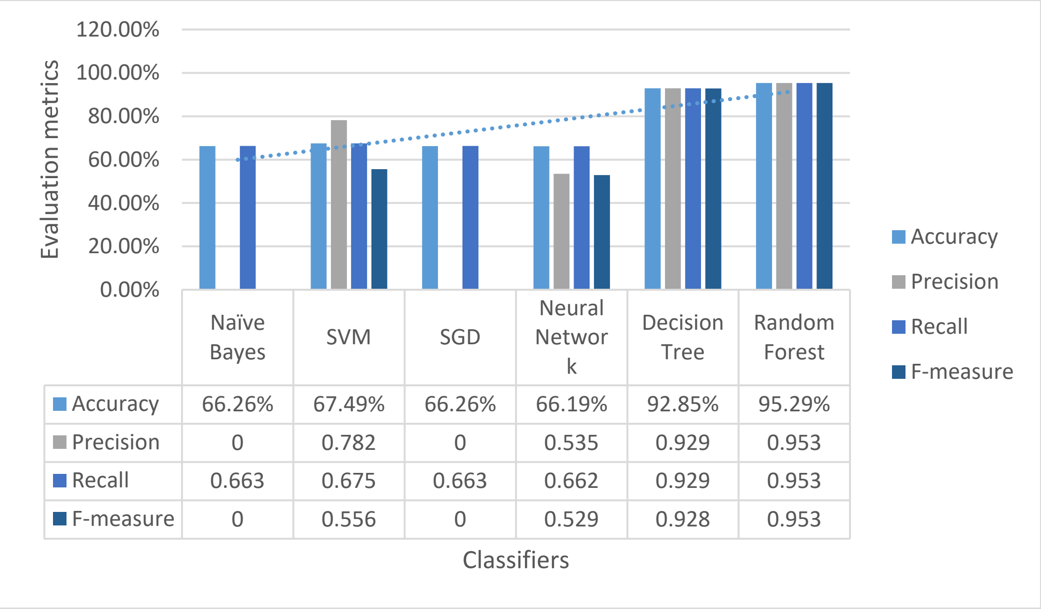
	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	68.77%	67.49%	68.72%	80.52%	93.37%	97.19%
Precision	0.682	0.782	0.765	0.821	0.933	0.972
Recall	0.688	0.675	0.687	0.805	0.934	0.972
F-measure	0.614	0.556	0.585	0.786	0.933	0.972

Figure 4.24 **Comparison of evaluation metrics for experiment 2 - Quality**

As it can be seen from the graphs above that Naïve Bayes performed at an average when the system was trained with all the categories in experiments 1 (Figure 4.16). Its accuracy suddenly increased by 10 points by training on two factors only from the Accuracy category and declined again on two factors of Authority category but close to previous category. (Figure 4.17 and 4.18). Aesthetics has only one factor and using it the classifier performed better than previous categories (Figure 4.19). The accuracies further divided lower than all the previous categories (Figure 4.20 and 4.21). This pattern shows that the misclassification ratio of Naïve Bayes is high as its performance is unreliable with many complex features. It performs better with less no. of factors in any category alone. The all categories together accuracy and the professionalism category accuracy, which has the most no of factors, is the lowest of all. This shows that Naïve Bayes does not work well with complex features. SVM has been consistent with no difference in performance with all or any single category.

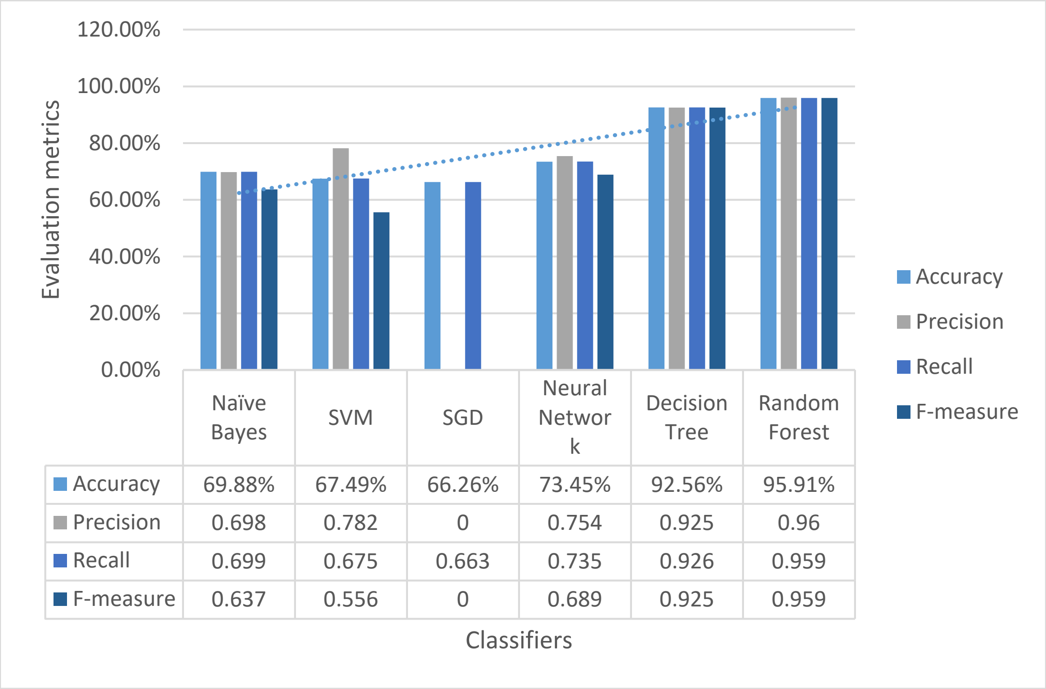
Neural Network has performed the best with all categories used together than any single category. Single categories' accuracies drastically drop with Neural Network. The three categories, professionalism (max features), popularity and quality with more features show an increase in performance while still being less than all categories together. This shows it reliability in tasks with many features. Decision Trees and Random Forest have consistently performed on top with accuracy above 90%. It can be clearly seen that Professionalism, Popularity and Quality categories with more no. of factors contribute more to credibility with better performance while Accuracy, Authority, Aesthetics and Currency with only one or two factors perform lower with an accuracy range of 92% - 95%. However, Random Forest performs remarkable in al given situations. This shows that Random Forest, with its ensemble property with Decision Trees, is most suitable for problems where accuracy matters.

Accuracy comparison graphs for category combinations



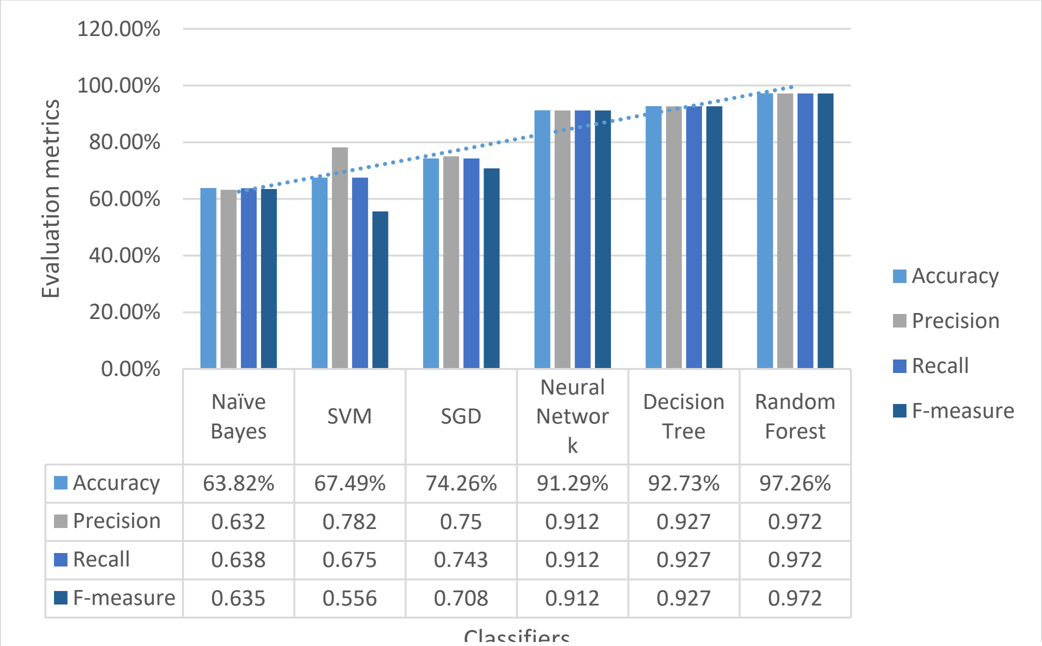
	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	66.26%	67.49%	66.26%	66.19%	92.85%	95.29%
Precision	0	0.782	0	0.535	0.929	0.953
Recall	0.663	0.675	0.663	0.662	0.929	0.953
F-measure	0	0.556	0	0.529	0.928	0.953

Figure 4.25 **Comparison of evaluation metrics for experiment 3 - combination 1 categories (Accuracy and Authority)**



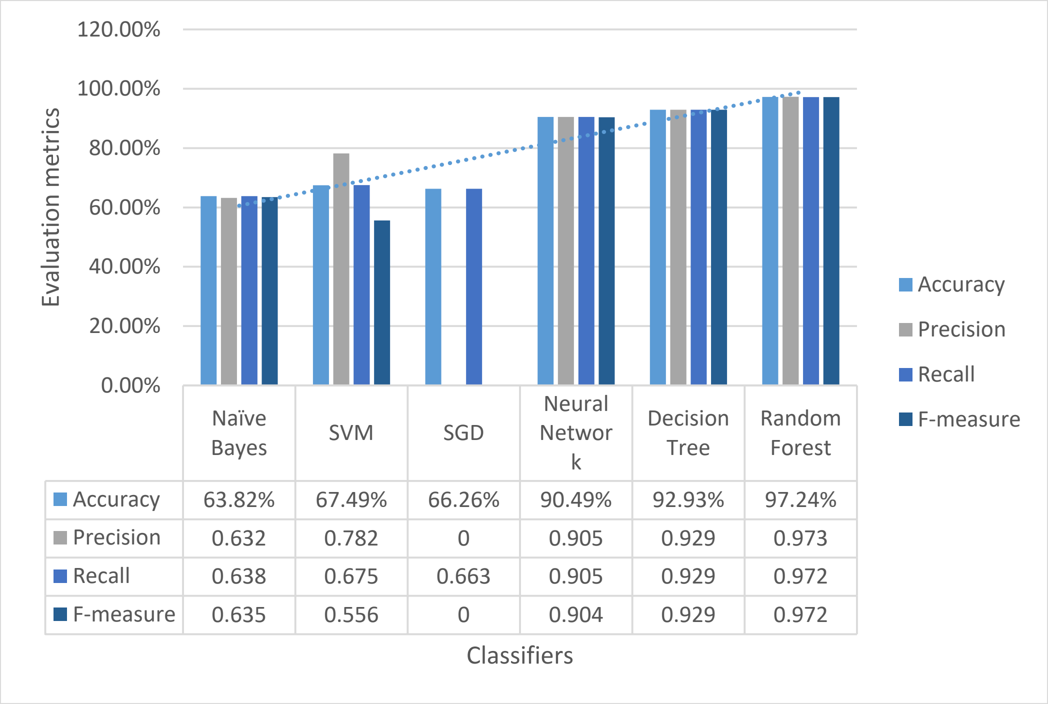
	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	69.88%	67.49%	66.26%	73.45%	92.56%	95.91%
Precision	0.698	0.782	0	0.754	0.925	0.96
Recall	0.699	0.675	0.663	0.735	0.926	0.959
F-measure	0.637	0.556	0	0.689	0.925	0.959

Figure 4.26 **Comparison of evaluation metrics for experiment 3 - combination 2 categories (Accuracy, Authority and Aesthetics)**



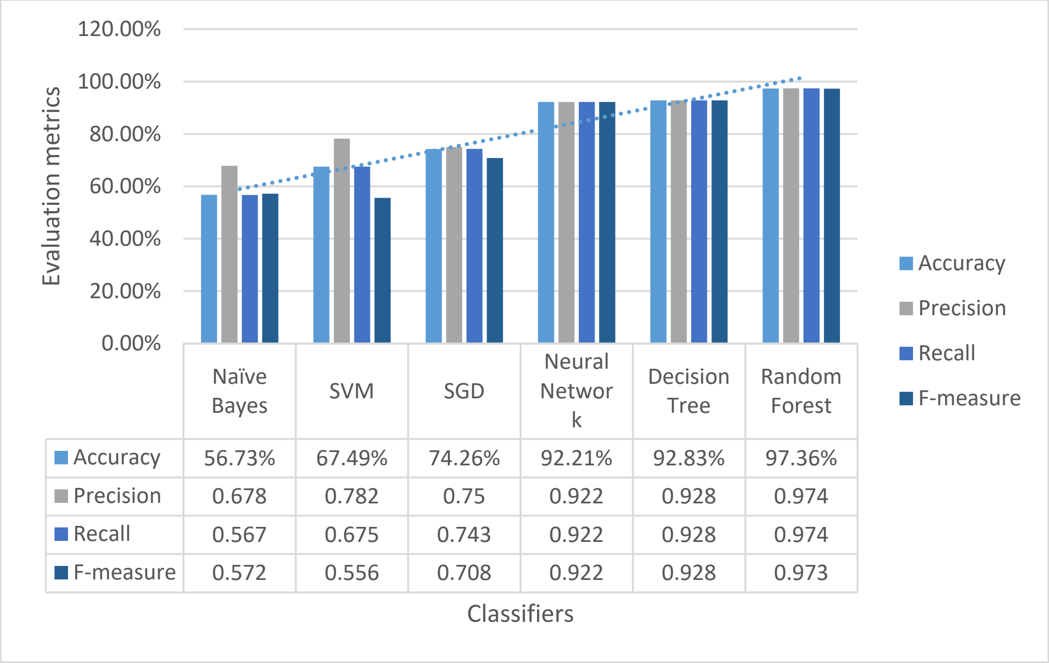
	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	63.82%	67.49%	74.26%	91.29%	92.73%	97.26%
Precision	0.632	0.782	0.75	0.912	0.927	0.972
Recall	0.638	0.675	0.743	0.912	0.927	0.972
F-measure	0.635	0.556	0.708	0.912	0.927	0.972

Figure 4.27 **Comparison of evaluation metrics for experiment 3 - combination 3 categories (Accuracy, Authority, Aesthetics and Professionalism)**



	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	63.82%	67.49%	66.26%	90.49%	92.93%	97.24%
Precision	0.632	0.782	0	0.905	0.929	0.973
Recall	0.638	0.675	0.663	0.905	0.929	0.972
F-measure	0.635	0.556	0	0.904	0.929	0.972

Figure 4.29 **Comparison of evaluation metrics for experiment 3 - combination 4 categories (Accuracy, Authority, Aesthetics, Professionalism and Popularity)**



	Naïve Bayes	SVM	SGD	Neural Network	Decision Tree	Random Forest
Accuracy	56.73%	67.49%	74.26%	92.21%	92.83%	97.36%
Precision	0.678	0.782	0.75	0.922	0.928	0.974
Recall	0.567	0.675	0.743	0.922	0.928	0.974
F-measure	0.572	0.556	0.708	0.922	0.928	0.973

Figure 4.28 **Comparison of evaluation metrics for experiment 3 - combination 5 categories (Accuracy, Authority, Aesthetics, Professionalism, Popularity and Currency)**

As shown in Figure 4.24 and 4.25, Naïve Bayes performs the best with category 1 and category 2 combination. After adding the Professionalism, Popularity and Quality, the accuracy of Naïve Bayes decreases than before. This shows again that Naïve Bayes does not perform well with more no. of factors/features. SVM has remained consistent in its performance in all experiments and combinations. Neural Network's accuracy increases as Professionalism, Popularity and Quality i.e., categories with more no. of factors are added (Refer to Figure 4.26, 2.27 and 2.28). With the combination of categories, there has been no major difference in performance of Decision Tree, however, the overall performance is greater than any other classifier except Random Forest. It remains at 92% with some difference in decimal points. Finally, Random Forest again outperforms all other classifiers with drastic increase in accuracy. With the addition of the bigger three categories, its accuracy is even better at 97%. This pattern shows the trends of classifiers such as Naïve Bayes performs better with less no. of features and performance remains average. While Neural Network and Random Forest perform better with complex features with higher accuracy.

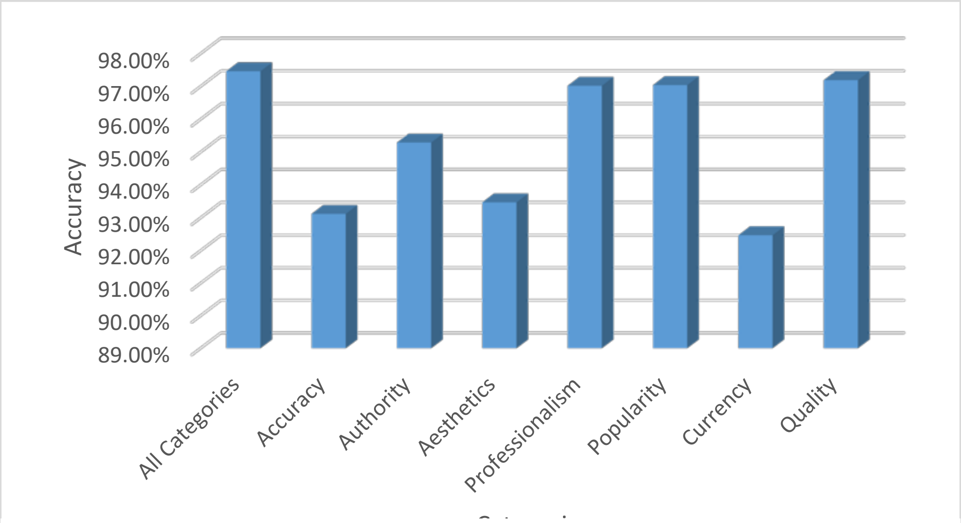
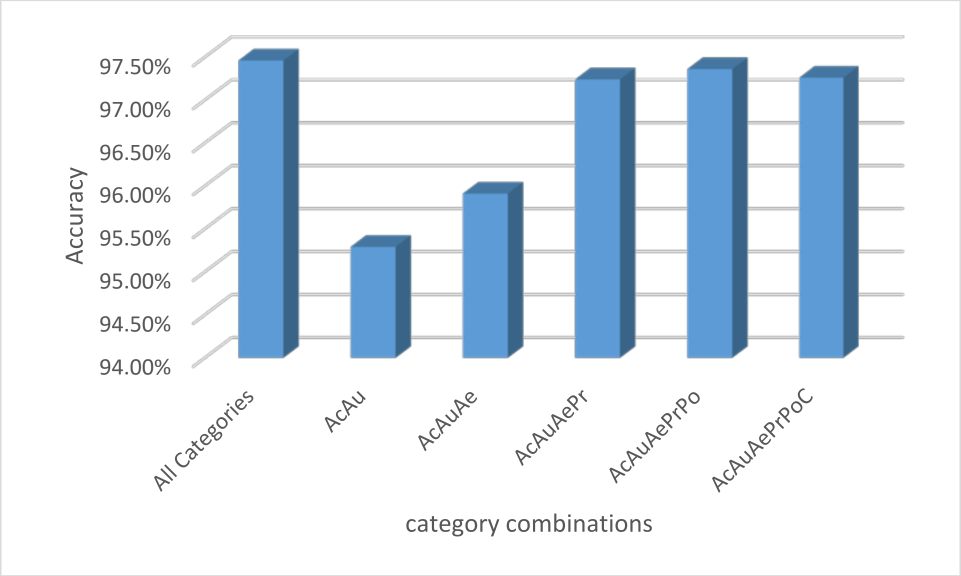


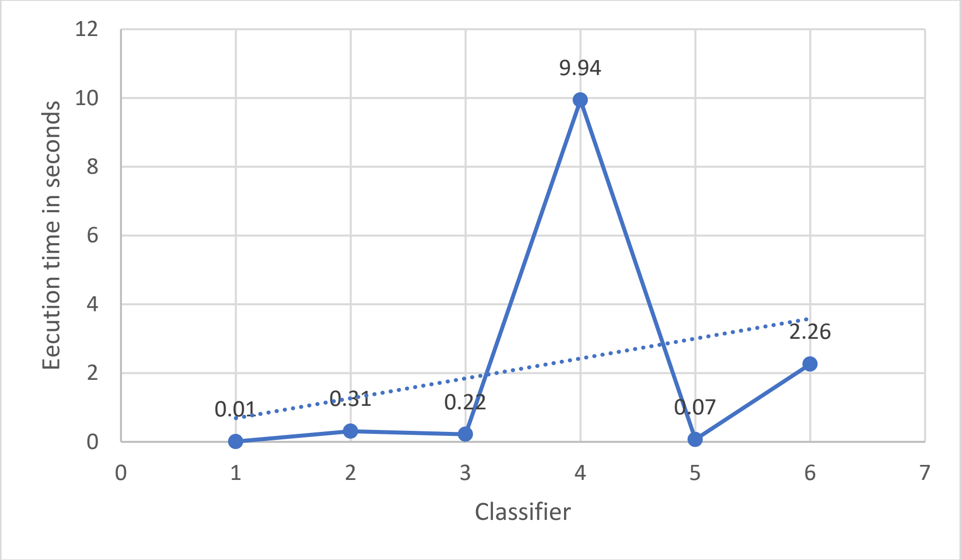
Figure 4.30 **Comparison of Random Forest accuracy between single categories and all categories**



Ac = Accuracy Au = Authority Ae = Aesthetics
Pr = Professionalism Po = Popularity C = Currency

Figure 4.31 **Comparison of Random Forest accuracy between combination of categories and all categories**

4.2.2. Result discussion for experiments based on execution time and accuracy



1 = Naïve Bayes 2 = SVM 3 = SGD
4 = Neural Network 5 = Decision Tree 6 = Random Forest

Figure 4.32 **Execution time (in seconds) graph of classifiers - all categories**

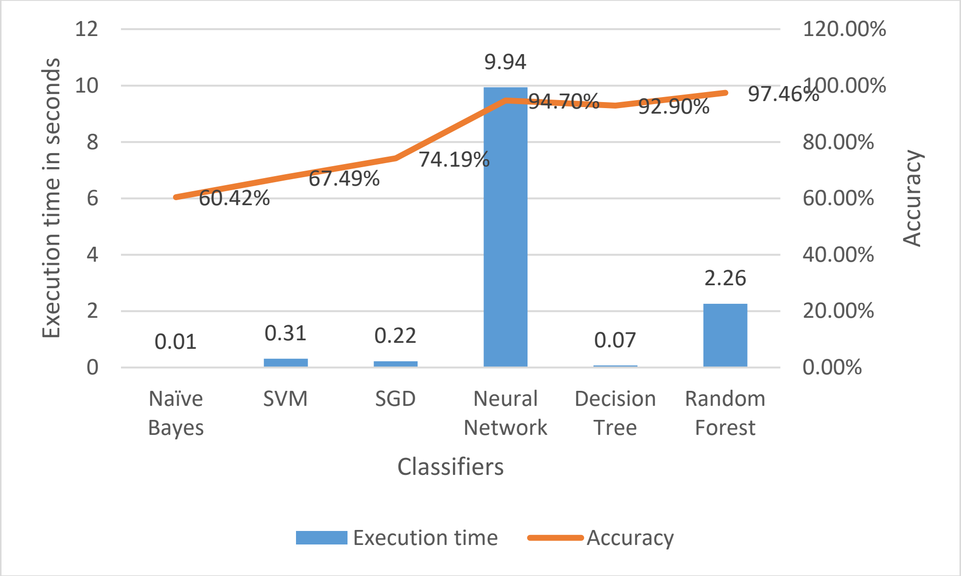
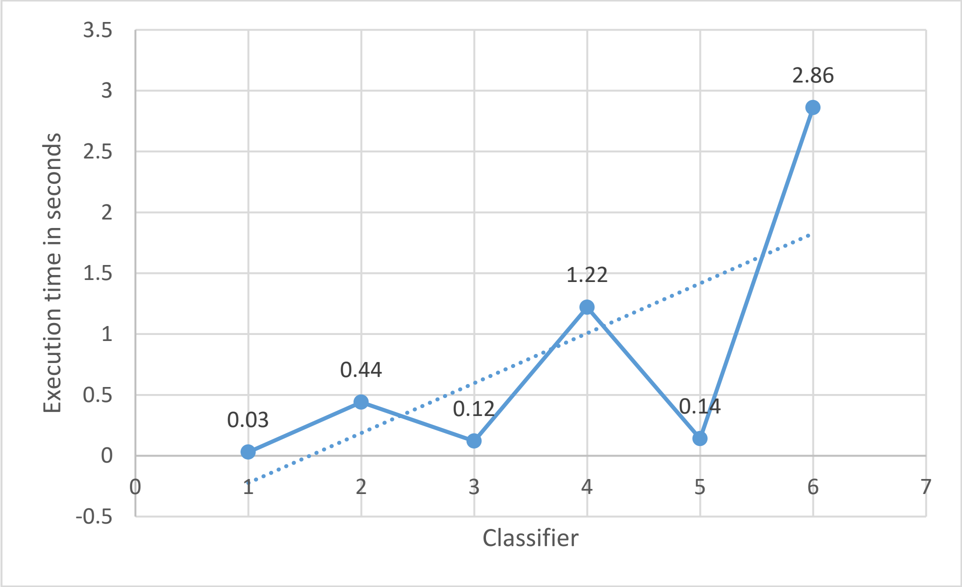


Figure 4.33 **Execution time (in seconds) vs accuracy of all classifiers – all categories**

The misclassification ratio of Naïve Bayes is high as its performance is unreliable with many complex features. However, the execution time of Naïve Bayes is the lowest as shown in figure 4.32. SVM and SGD still perform average with an increased execution time making them unsuitable for this problem. Decision Tree has a remarkable execution time with a great increase in accuracy. Neural Network and Random Forest are the top 2 performing classifiers. However, the difference in their execution time is huge making Neural Network computationally very expensive for an accuracy that is closer to the best. For a problem where *better computational complexity* with *good accuracy* is required, Decision Tree would be the best option.

Naïve Bayes has used the least no. of resources and is lightweight giving the best execution time than other classifiers but the trade-off between the execution time and accuracy for Naïve Bayes and Random Forest is not comparable. The linear dotted line in figure 4.31. shows the execution time trend of the classifier where Neural Network is an outlier. Based on these metrics, we choose Random Forest as the classifier for our methodology with the best accuracy and good efficiency.



1 = Naïve Bayes   2 = SVM   3 = SGD
4 = Neural Network   5 = Decision Tree   6 = Random Forest

Figure 4.34 **Execution time (in seconds) graph of classifiers - Authority**

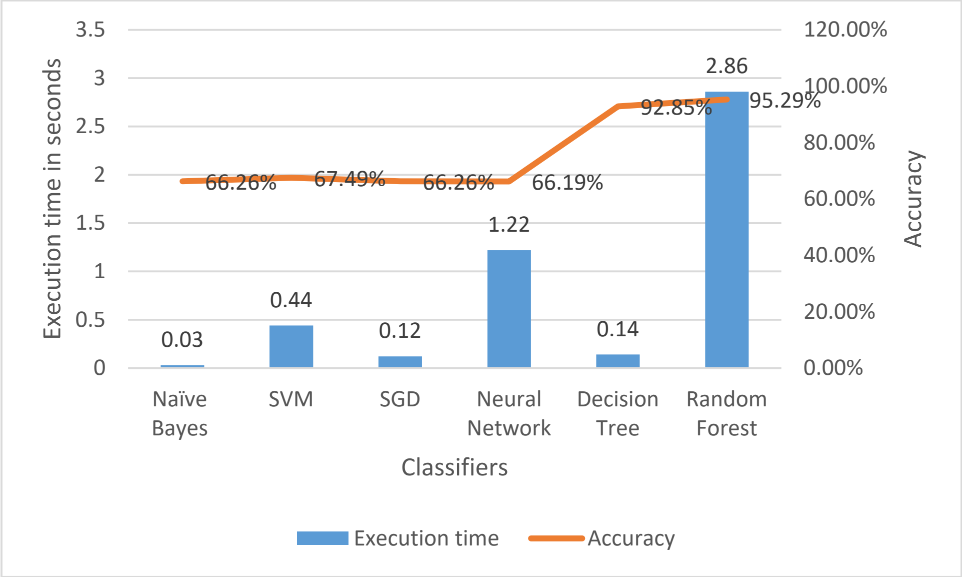
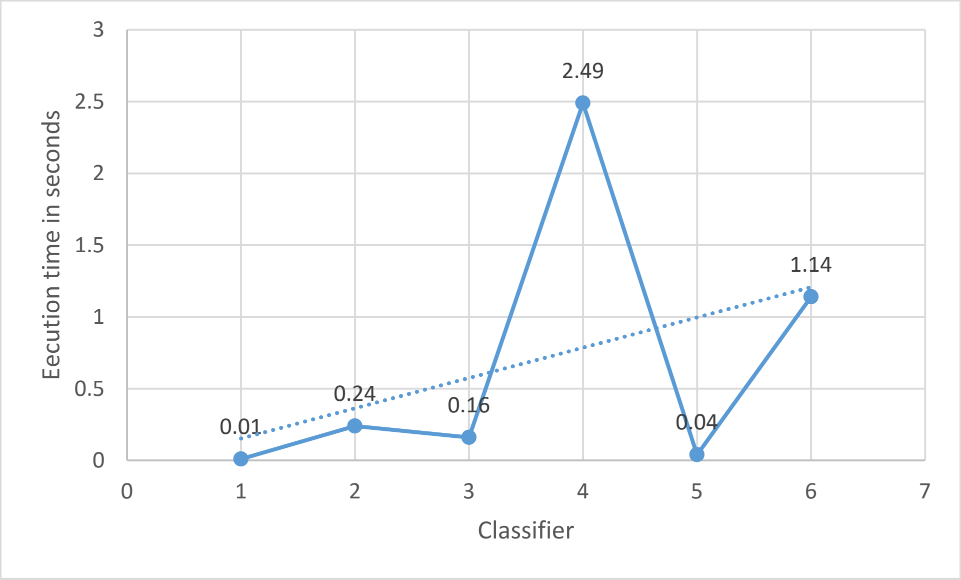


Figure 4.35 **Execution time (in seconds) vs accuracy of all classifiers – Authority**

As it can be seen in figure 4.34, Naïve Bayes and SVM perform very poorly with this category as their accuracies are average but execution time has increased using only 2 factors as compared to experiment 1 with 24 factors (figure 4.32). However, Neural Network's execution time has drastically decreased making it computationally cheap but the accuracy has also decreased a good amount. The execution time for Decision Tree has doubled with no increase in accuracy. As for Random Forest, it again performs the best with highest accuracy while its execution time remains similar to experiment 1. Though the accuracy is still less than collective categories.



1 = Naïve Bayes 2 = SVM 3 = SGD
4 = Neural Network 5 = Decision Tree 6 = Random Forest

Figure 4.36 **Execution time (in seconds) graph of classifiers - Professionalism**

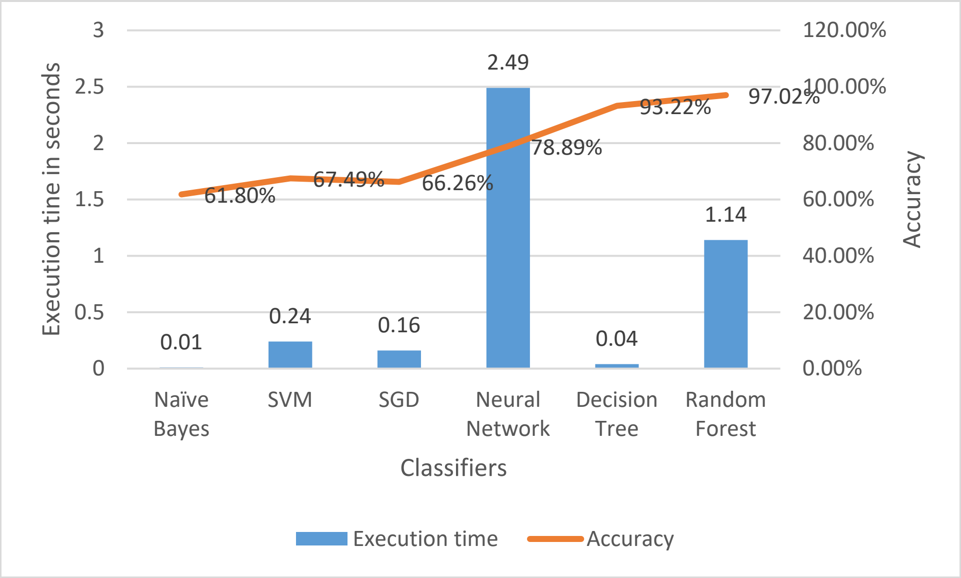


Figure 4.37 **Execution time (in seconds) vs accuracy of all classifiers – Professionalism**

As it can be seen from figure 4.36, with increasing the no. of factors compared to Authority category, Naïve Bayes again gives the lowest execution time (which is exactly same as experiment 1 with all categories) with an average accuracy (which is almost same as or with a minor increase in accuracy than experiment 1). SVM and SGD show no difference in accuracy as compared to Authority category but execution time decreases for SVM while it increases for SGD than experiment 1. Neural Network accuracy has decreased a lot but at the same time its execution time has lowered a great no. SVM shows no difference in accuracy with all, less or more factors but the execution time differs in each scenario. It decreases than all and less factors of Authority category.

4.2.3. Contribution of Aesthetics category

Certain studies [22], [31], [32] and [44] argued in favour of use of high aesthetics treatment websites for website credibility. They proposed that an aesthetically pleasing website adds to credibility incrementally. While certain studies [26] and [46] argued against use of aesthetics (section 2.1.2.3) and suggested that web templates can mask websites. Therefore, this study experimentally proved its contribution.

Lastly, these experiments show that a website with good aesthetics contribute to credibility collectively with other factors. An accuracy of 93.47% was achieved by Random Forest when the system was trained using Aesthetics alone. Eliminating this category will reduce the credibility. This category is rarely recognized as important. This study has proved its contribution experimentally.

4.3. Comparison with the baseline

The baseline method [28] from which the dataset was obtained uses a single feature for finding a reliable or trustworthy website to resolve conflicts. The feature used by them is data richness/domain expertise that can be mapped to authority category. The method used by them is statistical with Bayesian analysis. In comparison, this study works with 24 different features belonging to 7 categories for finding a credible website. They achieved an *accuracy of 87.77%, Precision of 0.877, Recall of 0.932 and F-measure 0.904* With incorporating machine learning, Random Forest proved to be performing well in all the situations with an *accuracy of 97.46%, precision of 0.975, Recall of 0.975 and F-measure 0.974* and hence, it was selected for our methodology. With 4000 rows, their execution time was *1.155s* while execution time of Random Forest with all the categories was *2.26s*. However, this is due to the ensemble property of Random Forest where the training time is more as it works with a forest of Decision Trees. Also, the no. of factors used by our system is large (24).

CHAPTER 5

Conclusion and future work

5. Conclusion

This papers aims at discovering true value from multiple conflicting sources for the same object and designs an approach for data fusion and web credibility using machine learning techniques. The proposed approach utilizes seven categories of web credibility that are discovered from various studies namely; Accuracy, Authority, Aesthetics, Professionalism, Popularity, Currency and Quality. After applying feature reduction techniques to 100s of identified factors for each category, suitable factors for the dataset contributing to each of these categories are recognized. Their values are derived using multiple APIs and are transformed to numerical form if required. A predictive model is built to evaluate web credibility based on the Book-Author dataset. Six supervised learning algorithms namely; Naïve Bayes, Support Vector Machine, Stochastic Gradient Descent, Neural Networks, Decision Trees and Random Forest are used. A series of different experiments are performed using these classifiers. Firstly, all categories are used together concretely to train all these classifiers. Secondly, each category is individually used and thirdly, a combination of categories is used to see their contribution in web credibility. Different evaluation metrics used for classifier performances based on these experiments revealed that Random Forest performs remarkable and is best fitted for predicting web credibility. The experiments proved that each category contributes to credibility and collectively these provide the best results. When used separately and in combination, three categories namely Professionalism, Popularity and Quality contributed more than the rest. Accuracy, Authority, Aesthetics and Currency contain only a maximum of two factors while Professionalism, Popularity and Quality contain many factors that contribute to their better performance. The former three achieved accuracies of 97.02%, 97.04% and 97.19% when used separately. But none of these categories separately performed as good as collectively. This study also identified the contribution of aesthetics in web credibility experimentally which is rarely recognized. It provided an accuracy of 93.47% alone which shows it’s important in credible websites. Our approach achieves a significant higher accuracy of 97.45% than the baseline accuracy *87.77%* reported by the authors by using all the 7 categories. However, the execution time with Random Forest is 2.26s which is greater than the baseline with an execution time of 1.15s. This is due to large no. of factors used in our system. Also, the trade-off between execution time and accuracy between the two methods is incomparable.

5.1. Future work

This study is aimed to be a subsequent contribution in an attempt to provide an automated solution to the truth finding problems. Hence, estimating source credibility and providing true answers supported by credible sources was the vital challenge in this study. The proposed solution provides significant solution with important credibility factors with great accuracy. But poses a limitation of the computational time of the classifiers.

This work can be further extended by adding more web content instances and dimensions and repeating the experiments with other advanced techniques to implement a more accurate and efficient system. Importance of credibility categories and heir subsequent factors may vary over time. Thus, highlighting and including other important factors over time on increased size of data will open new challenges for truth finding problems. In addition, multiple truths discovery and copying mechanisms can be tested further in this methodology.

References

[1] M. Z. Haroon, Z. Zeb, Z. Javed, Z. Awan, Z. Aftab, and W. Talat, “Internet Addiction In Medical Students,” *J. Ayub Med. Coll. Abbottabad*, vol. 30, no. 4, pp. S659–S663, 2018.

[2] L. Lenert and S. Skoczen, “The internet as a research tool: Worth the price of admission?,” *Ann. Behav. Med.*, vol. 24, no. 4, pp. 251–256, 2002, doi: 10.1207/S15324796ABM2404_01.

[3] I. World Stats, “Internet World Stats: Usage and Population Statistics,” 2019. https://www.internetworldstats.com/stats.htm%0Ahttps://www.internetworldstats.com/stats.htm%0Ahttps://www.internetworldstats.com/stats.htm%0Ahttps://www.internetworldstats.com/stats.htm%0Ahttps://www.internetworldstats.com/stats.htm%0Ahttp://www.internetw

[4] Google, “Google Search Statistics - Internet Live Stats,” 2018. http://www.internetlivestats.com/google-search-statistics/

[5] Google, “Google’s Year in Search - Google Trends,” 2020. https://trends.google.com/trends/yis/2020/GLOBAL/

[6] X. Yin, J. Han, and P. S. Yu, “Truth discovery with multiple conflicting information providers on the Web,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 796–808, 2008, doi: 10.1109/TKDE.2007.190745.

[7] R. M. and T. W. L. Page, S. Brin, “The PageRank Citation Ranking: Bringing Order to the Web,” 1999. doi: 10.1109/IISWC.2012.6402911.

[8] T. T. Nguyen, T. C. Phan, Q. V. H. Nguyen, K. Aberer, and B. Stantic, “Maximal fusion of facts on the web with credibility guarantee,” *Inf. Fusion*, vol. 48, no. April 2018, pp. 55–66, 2019, doi: 10.1016/j.inffus.2018.07.009.

[9] L. Berti-Equille and J. Borge-Holthoefer, “Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics,” *IEEE Xplore*, 2015.

[10] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, “Truth finding on the deep web: Is the problem solved?,” *Proc. VLDB Endow.*, vol. 6, no. 2, pp. 97–108, 2012, doi:

10.14778/2535568.2448943.

[11] A. Galland, S. Abiteboul, A. Marian, and P. Senellart, “Corroborating information from disagreeing views,” *WSDM 2010 - Proc. 3rd ACM Int. Conf. Web Search Data Min.*, pp. 131–140, 2010, doi: 10.1145/1718487.1718504.

[12] X. L. Dong and F. Naumann, “Data fusion - Resolving data conflicts for integration,” *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1654–1655, 2009, doi: 10.14778/1687553.1687620.

[13] X. S. Fang, Q. Z. SHENG, W. Xianzhi, M. Barhamgi, and L. Yao, “SourceVote : Fusing multi-valued data via inter- source agreements,” in *Institutional Knowledge at Singapore Management University*, 2017, pp. 164–172.

[14] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” *Struct. Dyn. Networks*, vol. 9781400841, no. 5, pp. 514–542, 2011, doi: 10.1515/9781400841356.514.

[15] A. H. Ahmed and F. Sadri, “DataFusion – Taking source confidences into account,” in *ACM International Conference Proceeding Series*, 2018, pp. 1–6. doi: 10.1145/3200842.3200854.

[16] B. Fogg and S. Tseng, “Credibility and Computing Technology,” *Commun. Acm*, vol. 42, no. 5, pp. 39–44, 1999.

[17] S. Aggarwal, H. Van Oostendorp, Y. R. B. Reddy, and B. Indurkhya, “Providing Web Credibility Assessment Support,” in *European Conference on Cognitive Ergonomics (ECCE)*, 2014, no. September.

[18] A. A. Shah, S. D. Ravana, S. Hamid, and M. A. Ismail, “Web credibility assessment: Affecting factors and assessment techniques,” *Inf. Res.*, vol. 20, no. 1, pp. 1–28, 2015.

[19] Y. Li *et al.*, “A Survey on Truth Discovery,” *ACM SIGKDD Explor. Newsl.*, vol. 17, no. 2, pp. 1–16, 2016, doi: 10.1145/2897350.2897352.

[20] L. Berti-Équille, “Truth Discovery,” *Encycl. Big Data Technol.*, pp. 1718–1726, 2019, doi: 10.1007/978-3-319-77525-8_23.

[21] F. Luis and G. Moncayo, “SLFTD: A Subjective Logic Based Framework for Truth Discovery,” *Springer Nat. Switz. AG*, pp. 102–110, 2019.

[22] S. M. Shariff, “A Review on Credibility Perception of Online Information,” *Proc. 2020 14th Int. Conf. Ubiquitous Inf. Manag. Commun. IMCOM 2020*, 2020, doi: 10.1109/IMCOM48794.2020.9001724.

[23] H. Xiao *et al.*, “Towards Confidence Interval Estimation in Truth Discovery,” *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 3, pp. 575–588, 2019, doi: 10.1109/TKDE.2018.2837026.

[24] L. Berti-equille, “Truth Discovery: A Survey,” *Springer, Encycl. Big Data Technol.*, 2018.

[25] J. Pasternack and D. Roth, “Knowing what to believe (when you already know something),” *Coling 2010 - 23rd Int. Conf. Comput. Linguist. Proc. Conf.*, vol. 2, no. August, pp. 877–885, 2010.

[26] M. J. Metzger and A. J. Flanagin, “Credibility and trust of information in online environments: The use of cognitive heuristics,” *J. Pragmat.*, vol. 59, pp. 210–220, 2013, doi: 10.1016/j.pragma.2013.07.012.

[27] X. L. Dong, L. Berti-Equille, and D. Srivastava, “Integrating Conflicting Data: The Role of Source Dependence,” *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 550–561, 2009, doi: 10.14778/1687627.1687690.

[28] X. Lin and L. Chen, “Domain-aware multi-truth discovery from conflicting sources,” *Proc. VLDB Endow.*, vol. 11, no. 5, pp. 635–647, 2018, doi: 10.1145/3177732.3177739.

[29] F. Luis and G. Moncayo, “An Effective Truth Discovery Algorithm with Multi-source Sparse Data,” *Springer Int. Publ. AG*, pp. 434–442, 2018.

[30] D. Robins, J. Holmes, and M. Stansbury, “Consumer Health Information on theWeb: The Relationship of Visual Design and Perceptions of Credibility,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 61(1), no. July, pp. 13–29, 2010, doi: 10.1002/asi.

[31] T. Meng *et al.*, “Journal Pre-proof,” 2019.

[32] D. Robins and J. Holmes, “Aesthetics and credibility in web site design,” *Inf. Process. Manag.*, vol. 44, no. 1, pp. 386–399, 2008, doi: 10.1016/j.ipm.2007.02.003.

[33] V. A. Tsygankov, “Evaluation of website trustworthiness from customer perspective, a

framework," in *ACM International Conference Proceeding Series*, 2004, vol. 60, pp. 265–271. doi: 10.1145/1052220.1052254.

[34] R. Manjula and M. S. Vijaya, "Measuring Web Content Credibility Using Predictive Models," *Springer Nat. Singapore Pte Ltd*, vol. 89, no. Inventive Communication and Computational Technologies, pp. 21–38, 2020, doi: 10.1057/9781137379283_2.

[35] X. S. Fang, Q. Z. Sheng, X. Wang, D. Chu, and A. H. H. Ngu, *SmartVote: a full-fledged graph-based model for multi-valued truth discovery*, vol. 22, no. 4. 2019. doi: 10.1007/s11280-018-0629-3.

[36] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Truth discovery and copying detection in a dynamic world," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 562–573, 2009, doi: 10.14778/1687627.1687691.

[37] X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava, "Global detection of complex copying relationships between sources," *Proc. VLDB Endow.*, vol. 3, no. 1, pp. 1358–1369, 2010, doi: 10.14778/1920841.1921008.

[38] M. Wu and A. Marian, "A framework for corroborating answers from multiple web sources," *Inf. Syst.*, vol. 36, no. 2, pp. 431–449, 2011, doi: 10.1016/j.is.2010.08.008.

[39] C. Chang *et al.*, "An Unsupervised Approach of Truth Discovery from Multi-Sourced Text Data," *IEEE Access*, vol. 7, pp. 143479–143489, 2019, doi: 10.1109/ACCESS.2019.2934469.

[40] H. Zhang, Y. Li, F. Ma, J. Gao, and L. Su, "Text truth: An unsupervised approach to discover trustworthy information from multi-sourced text data," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 2729–2737, 2018, doi: 10.1145/3219819.3219977.

[41] N. Jnoub, W. Klas, P. Kalchgruber, and E. Momeni, "A Flexible Algorithmic Approach for Identifying ConflictingDeviating Data on the Web.pdf," in *International Conference on Computer, Information and Telecommunication Systems (CITS)*, 2018, pp. 1–5.

[42] S. Aggarwal and H. Van Oostendorp, "An Attempt to Automate the Process of Source Evaluation," in *Proc. of Int. Conf. on Advances in Computer Engineering*, 2011, vol. 02, no. 02, pp. 1–3.

[43] S. Ansari and J. Gadge, “Architecture for Checking Trustworthiness of Websites,” *Int. J. Comput. Appl.*, vol. 44, no. 14, pp. 22–26, 2012, doi: 10.5120/6332-8706.

[44] A. Ali Shah and S. Devi Ravana, “Evaluating information credibility of digital content using hybrid approach,” *Int. J. Inf. Syst. Eng.*, vol. 2, no. 1, pp. 92–99, 2014, [Online]. Available: https://www.researchgate.net/publication/320161434%0Afile:///C:/Users/Alexandre/Downloads/EvaluatingInformationCredibilityofDigitalContentusingHybridApproach.pdf

[45] H. Singal and S. Kohli, “Trust Necessitated through Metrics: Estimating the Trustworthiness of Websites,” *Procedia Comput. Sci. Int. Conf. Comput. Model. Secur.*, vol. 85, no. Cms, pp. 133–140, 2016, doi: 10.1016/j.procs.2016.05.199.

[46] J. Pattanaphanchai, K. O’Hara, and W. Hall, “Trustworthiness criteria for supporting users to assess the credibility of web information,” *WWW 2013 Companion - Proc. 22nd Int. Conf. World Wide Web*, pp. 1123–1130, 2013, doi: 10.1145/2487788.2488132.

[47] A. A. Shah, S. D. Ravana, S. Hamid, and M. A. Ismail, “Web Pages Credibility Scores for Improving Accuracy of Answers in Web-Based Question Answering Systems,” *IEEE Access*, vol. 8, pp. 141456–141471, 2020, doi: 10.1109/ACCESS.2020.3013411.

[48] Moz, “Domain Authority Checker - Moz,” 2022. https://moz.com/domain-analysis

[49] Moz, “Page Authority,” 2021. http://moz.com/learn/seo/page-authority

[50] dead link checker, “Broken or dead links,” *https://www.deadlinkchecker.com/website-dead-link-checker.asp*.