Development of a machine learning-based model to autonomously estimate web page credibility

by

Amir Mehdi Shamsi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2020

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

There is a broad range of information available on the Internet, some of which is considered to be more credible than others. People consider different credibility aspects while evaluating the credibility of a web page, however, many web users find it difficult to determine the credibility of all types of web pages. An autonomous system that can analyze different credibility factors extracted from a web page to estimate the page's credibility could help users to make better decisions about the perceived credibility of the web information.

This research investigated the applicability of several machine learning approaches to the evaluation of web page credibility. First, six credibility categories were identified from peer-reviewed literature. Then, their related credibility features were investigated and automatically extracted from the web page content, metadata, or external resources.

Three sets of features (i.e., automatically extracted credibility features, bag of words features, and combination of both) were used in classification experiments to compare their impact on the autonomous credibility estimation model performance. The Content Credibility Corpus (C3) dataset was used to develop and test the performance of the model developed in this research. XGBoost achieved the best weighted average F1 score for extracted features. In comparison, the Logistic Regression classifier had the best performance when bag of words features was used, and all features together were used as a feature vector.

To begin to explore the legitimacy of this approach, a crowdsourcing task was conducted to evaluate how the output of the proposed model aligns with the credibility ratings given by human annotators. Thirty web pages were selected from the C3 dataset to find out how current users' ratings correlate to the ratings that were used as ground truth to train the model. In addition, 30 new web pages were selected to explore how generalizable the algorithm is for classifying new web pages.

Participants were asked to rate the credibility of each web page base on a 5-point Likert scale. Sixty-nine crowd-sourced participants evaluated the credibility of the 60 web pages for a total of 600 ratings (10 per page). Spearman's ρ between average credibility scores given by participants and original scores in the C3 dataset indicates a moderate positive correlation: $\rho = 0.44, p < 0.02$. A contingency table was created to compare the predicted scores by the model with the rated scores by participants. Overall, the model achieved an accuracy of 80%, which indicates that the proposed model can generalize for new web pages.

The model outlined in this thesis outperformed the previous work by using a promising set of features that some of them were presented in this research for the first time.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Jennifer Boger, who always went above and beyond to help me during my master's studies. She was always available to offer her insightful thoughts and advice, and kindly supported me in all difficulties that I have encountered for the past two years.

I'm also thankful to my thesis readers, Dr. Alexander Wong and Dr. Helen Chen, for taking the time to read my thesis and providing their feedback.

I would like to thank all my friends at the ITWIL lab, especially Sheida and Jing, for making the lab a pleasant place to work. I also want to thank my lifelong friends, Keyvan and Nazanin, for their unconditional help and support.

Last, I want to thank my family, Mehrdad, Zahra, and my lovely niece, Parimah. I'm also grateful to my parents and my sister, Mojgan, whom I really missed. Thank you for all your love and support.

The research reported in this thesis was funded by the AGE-WELL NCE Inc. as part of the WP2.3 CARE-RATE project.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thesis Motivation

Since its introduction, the Internet has become an increasingly important source of information that many people rely on for decision making in their daily lives [5, 75]. People refer to websites to obtain all sorts of information, such as news, financial information, and medical information [102]. The amount of available information on the web is roughly doubling every two years, however, most of the information is not rigorously fact-checked before publishing [75, 94]. Consequently, due to the open and unmonitored nature of the web, there exist numerous web pages on the Internet that contain fake, incorrect, or misleading information [4, 75, 82, 102]. For instance, while lots of users, including patients and caregivers, search for health information on the web, there has been no quality control by medical specialists for more than half of the existing medical websites on the Internet [86, 87, 100].

The term 'information credibility' means how believable a piece of information is perceived by the person accessing it [30, 67, 88]. A web page that is perceived to be credible is defined as a web page *"whose information one can accept as the truth without needing to look elsewhere"* [82]. In other words, people believe a web page to be credible if it is one that the viewer perceives as being trustworthy and having high quality and accurate content [26]. These definitions are the most widely used definitions for the web credibility in past research and are used in this thesis as well.

Many researchers consider credibility as a perceived quality that consists of both objective and subjective components and is different from the truth, which is usually considered to be more of an objective quality [93, 96]. The subjective part is associated with users' perception, whereas the objective part is relevant to source or content attributes [24]. This is why the credibility of information can be interpreted differently depending the type of information and on who the person evaluating it is [30, 68]. Assessing several information aspects at the same time forms each user's perception of credibility [26].

An important impact of information credibility is the fact that it affects perceived usefulness and perceived risk of a website; positive feelings about website credibility leads to users' inclination to trust the information and to return to that website again [64]. With the increasing pervasiveness of the Internet and the amount of information on it, finding credible information is of growing importance. Having an estimate of the objective aspects of credibility could support people in making better, quicker decisions regarding web page credibility. This could be helpful for many Internet users, particularly those whom may not be familiar with the information they are looking at or are less experienced in Internet searching (e.g., many older adult caregivers).

1.2 Research Questions

The research presented in this thesis was guided by the following research questions:

1. What features and methods could be used to measure different aspects related to web page credibility?

2. What model can be developed to automatically ascertain web page credibility?

3. How does the estimated credibility level by the model align with the perceived credibility of human web users?

1.3 Thesis Organization

The rest of the thesis is organized as follows:

Chapter 2 focuses on peer-reviewed works related to the credibility assessment of web pages. It discusses several approaches to evaluate the credibility of different types of online resources automatically. Chapter 3 describes various credibility factors from different credibility categories that are used to build the model. It also explains the methods that are used in this research to extract credibility features from web pages automatically. In Chapter 4, the dataset that was used to train the model is described and the machine learning algorithms to create the model, evaluation metrics, and the classification and regression results from conducting different machine learning experiments are discussed. Chapter 5 presents the experiment conducted to explore the generalizability of the model for new web pages outside of the training dataset, where crowdsourced ratings are compared to the output of the model to evaluate its ability to estimate perceived credibility. Chapter 6 summarizes the key findings and conclusions from this research as well as possible future work in the autonomous assessment of web page credibility.

Chapter 2

Background

2.1 Assessing Credibility

In order to create a model that estimates the level of the credibility of a web page, it is essential to understand what factors impact the perceived credibility of web pages and how users use those factors to evaluate the credibility of web pages. Credibility factors can be described as any of the features of the resource that contribute to the credibility of a web page, including information content characteristics, information source details, or other external features. A combination of the related factors creates a framework to assess a specific aspect of credibility. Different aspects, such as expertise, accuracy, and quality, are called credibility categories [84]. Each discipline emphasizes certain aspects of credibility more than others. For example, content accuracy and quality are considered more important in information science, while information related to psychology and communication has more of an emphasis on the source reliability and reputation [30].

Trustworthiness and expertise are two major credibility components that are used in credibility assessment by users. The trustworthiness component relates to the perceived goodness of the source and can be described by terms such as unbiased, truthful, and well-intentioned. On the other hand, the expertise component refers to the perceived knowledge and skill of the source. Terms such as experienced, knowledgeable, and competent can describe the expertise credibility dimension [93]. The Prominence-Interpretation theory proposed by [28] assumes that two things occur when users evaluate the credibility of a web page; first, users must notice a feature or element of the page (Prominence) and then they can make a judgment about its credibility (Interpretation). Prominence is affected by factors such as the topic of the web page, involvement, experience, individual differences, and the task of the user. Interpretation and user's judgment about an element are impacted by the user's assumptions in mind, the user's skill or knowledge, and context such as the user's environment or expectations [28].

People typically use heuristics and their prior expectations to evaluate credibility [73, 80]. Many users consider search result ranking position of web pages as an essential indicator of web page credibility, which is not always correct [43, 77]. When users browse an unfamiliar website, the initial impressions greatly influence their perception of its credibility, including now the site looks. The reason these judgements are made this way is because when people lack prior knowledge of the web page, and there is no other information available, they usually trust their initial feelings, which are often significantly affected by information source appearance [35, 64].

In general, both information content and source characteristics significantly influence users' perception of the credibility of a web page [37]. The content author's expertise and credentials in addition to the domain type of the website are some of the source characteristics that play an important role in users' credibility judgment of the website [66, 67]. Users' time restrictions and significance of the information to them (e.g., if they are looking for information about a critical health problem) can considerably impact their credibility assessment as well [6, 37].

Several studies have investigated key factors during credibility assessment. Eighteen areas that users regularly notice in credibility evaluation were outlined by [29]. This research found visual design, information structure, and information focus of the site as mostly mentioned factors by study participants in evaluating the credibility of web pages. [26] developed seven scales to explore how different aspects impact website credibility. Real-world feel (i.e., aspects that map to an organization's real-world presence, such as location and contact information), trustworthiness, ease of use, expertise, and tailoring were found to have positive effects on users' perception of credibility. For example, when a when a website provides a quick response to customer service requests, its real-world feel is improved. Conversely, amateurism and commercial implications (i.e., containing advertising or asking for a payment) had a negative effect on credibility.

Three types of features are suggested by [62] that are important in assessing the credibility of a web page: 1) semantic features such as accuracy or sentiment, which are more influential if users are domain experts; 2) surface features like web page design, page color, and page layout; and 3) users' previous experience with the information source. Credibility features can also be categorized into two main classes: on-page features and off-page features. On-page features are features that exist on the web page; however, users need to take time to pay attention to them and it might not be easy for them to recognize or quantify these features. Domain type and spelling errors are some examples of this type of features. On the other hand, some features are extracted from metadata or users are required to look somewhere other than the target web page to obtain them. These features, such as page ranking or Google index, are called off-page features and are harder for the user to ascertain [82].

Web page credibility assessment can be performed by humans (users) or by computers. Since no web user has the experience and skills to determine the credibility of all types of web pages, it would be helpful to provide tools that can support assessment of web information credibility. This is especially true in critical areas such as financial, health, and medical information [4, 24, 69, 75]. Using or trusting unreliable information in these areas may cause severe consequences for many types of users [87, 91]. One problem with relying on humans as credibility evaluators is that the evaluation outcome is usually inconsistent due to individual differences and discrepancies in users' perceptions and interpretation of information. In addition, these methods are generally very time-consuming and require a lot of effort, training, and motivation by users. Some alternative approaches to human credibility evaluation are credibility rating systems, digital signatures, and collaborative filtering [67, 84].

While many web users accept online information without checking its credibility [100], providing them with an estimated credibility score of online information was reported to be helpful for them and increased their confidence about that information especially in case of inexperienced users [4, 82]. Past studies reveal that even presenting simple information such as contact information or details about the author or publisher of the page can lead to a considerable enhancement in users' credibility perception of a web page [32, 84]. Another critical aspect is that providing information about the credibility of a website can also have positive effects on the success of the website. The reason is that if users deem a web page as non-credible, it is unlikely that they continue using its products or services anymore [28]. Users' positive credibility perception of a web page leads to more trust in that web page, which helps to decrease uncertainty about the information [37, 59].

In brief, there is a broad range of online information available on the Internet with different levels of credibility. Therefore, using an autonomous system that can collect, measure, and analyze different credibility factors extracted from a web page to estimate the page's credibility could help users to make faster and better decisions about the trustworthiness of information they encounter on the web. These types of automated systems could be most useful for inexperienced users who are more exposed to misleading online information [7, 73].

2.2 Previous Works on Assessing Credibility

Different approaches have been used by researchers to investigate web credibility from various perspectives. Many studies investigated how people assess credibility to provide web designers with credibility guidelines that need to be considered in order to improve the credibility perception of users [26, 85]. Some researchers examined the credibility of specific subsets of online information such as weblogs [50, 80] and social media platforms like twitter [12, 40, 41, 70]. While several studies investigated the quality of content as a significant indicator of credibility [5, 49, 72], some studies focused on detecting spam web pages and separating them from credible ones [13, 42]. Also, a number of studies aimed to inform search engine users about various credibility-related features of returned results using visual cues [7, 82, 100, 102]. Different automated systems have been developed that make use of a combination of content and social features to evaluate the credibility of web pages [4, 75, 87, 91, 96]. The rest of this chapter focuses on approaches and features used in prior research for computer-based assessment of website credibility.

2.2.1 Assessing Credibility of Social Media

While some autonomous credibility assessment methods are suitable for evaluating the credibility of any general web page, some methods are specifically designed for certain types of online information. Social media and user-generated content are becoming more popular on the web. Thus, several studies have been conducted to evaluate the credibility of such subsets of the web, including weblogs, web forums, question-answering portals, as well as social networking and microblogging platforms like twitter [70, 40, 12, 41]. While these platforms might have their own specific features for the credibility assessment that are not available on regular web pages, some of the methods for extracting content features are still applicable to other platforms, such as word or sentiment features. In credibility research associated with weblogs, [50] developed a blog credibility classification model based on various content, sentiment, and style features identified from cognitive procedures used by human evaluators. In another weblog related research, [80] created a framework to assess the credibility of weblogs based on bloggers' expertise and trustworthiness, quality of the information, and users' personal preferences. They identified and ordered credibility factors considering their perceived importance to users.

Unlike web pages, visual design is unrelated to tweets credibility perception. [70] demonstrated that users' perceived credibility of tweets is not solely dependent on tweet content, instead it is mostly affected by the source characteristics such as the number of followers, mentions, and topical expertise of the tweet author. Their results indicated that the author's username and profile image could influence the credibility assessment of tweet readers no matter the actual content of the tweet is trustworthy or not. Information credibility of tweets regarding high impact news events through identifying various content and source-based features was investigated by [40]. They found the number of followers and username length as the most impactful source features. The number of unique characters, pronouns, swear words, and emoticons in a tweet observed to be the most important content features. They re-ranked the tweets about an event based on their predicted credibility score using Pseudo Relevance Feedback technique. Also, [12] proposed a method for automatic credibility evaluation of a particular set of tweets about a trending topic. They used features such as users' tweeting and retweeting behavior, tweet text and external source references to create their model. Another method to automatically evaluate the credibility of Twitter events is suggested by [41]. First, they perform a credibility propagation on a network composed of tweets, events, and users. For credibility propagation, they used iterations similar to PageRank [76] algorithm. Then, by assuming that the credibility score of similar events should be similar, they created another graph of events in each iteration to optimize the credibility scores.

Content quality is deemed to be a prominent indicator of web credibility. [5] investigated social media elements that help find high-quality content, particularly in online question-answering portals like Yahoo! Answers. They used both syntactic and semantic complexity features as well as formality scores to approximate the grammatical quality of the content. In addition to the content features, they modeled users' interactions and feedback in a graph-based framework. Topic coverage and topic detailedness are two query-dependent measures proposed by [72] to assess the quality of web pages. Topic coverage means how many typical topics relevant to the query are covered by the web page, while topic detailedness counts the number of special topics on the page. These two measures depend on the expertise of users as well because expert users are likely to prefer more specific web pages related to their queries. In contrast, novice users usually favor general topics associated with their queries. Their proposed system initially identified the domain of the search query, then typicality and specialty scores were computed by analyzing the link structure of the related pages in Wikipedia. The problem with this approach is that Wikipedia articles may not contain all the queried topics with appropriate quality and scope [72]. Many online service providers keep track of the number of users clicks or recommendations, [49] has used such non-textual features to create a framework to estimate quality and consequently credibility of answers in a web-based question-answering platform.

2.2.2 Credibility Rating Systems

Credibility rating systems and collaborative filtering systems rely on users' and experts' ratings to evaluate the credibility of online information. For instance, Web of Trust (mywot.com) is a website reputation and review service that uses worldwide crowdsourcing to collect comments and ratings from millions of users regarding the trustworthiness of websites [51]. If a group of people has a general agreement on the trustworthiness of some information, it is more likely that the information is deemed correct and credible by other individuals [37]. Social consensus can have a significant impact on the acceptance of particular information because it can decrease uncertainty and perceived risk of using that information [11, 92]. There are different types of online social feedback systems, including rating systems, recommender systems, and social navigation tools [37]. The effect of social feedback, like audience ratings on credibility perception, has been examined by [37]. The results showed that while the type of feedback such as positive or negative does affect users' perceived credibility, the size of the audience does not seem to be that influential. They also observed that users with prior knowledge about the topic of a web page are less likely to consider audience feedback when assessing credibility.

[77] proposed a social recommender system to evaluate the credibility of web pages based on collaborative filtering. Their system is composed of three major components to approximate web page credibility: 1) a social component that makes use of credibility ratings given by user's friends, 2) a content component that employs content-based features like semantic, syntactic, and sentiment ones, and 3) a ranking component that uses ranking of the page in search results. By combining these three components using adaptive weights, they re-rank the search results provided by a search engine for users.

It is worth noting that while aggregate user opinions are helpful in assessing credibility of a web page, it is not as objective or consistent as a machine-based approach could be. That said, an autonomous assessment must return results that reflect what humans consider to represent credible content.

2.2.3 Augmenting Credibility of Search Engine Results

Many web users acquire the information they are looking for through search engines. Search engines provide a list of relevant results to users' queries without validating the credibility of the source that most users either consider them as credible or they need to assess the credibility of results themselves [99, 101]. However, current search engines display few web page features such as titles, snippets, and URLs that are not enough for evaluation of credibility appropriately [102]. Some studies have been aimed at helping users have better credibility assessment by automatically identifying false facts on the page or by providing users with information that improves their perception of the quality of the page and enhances information transparency [82]. While visualizing credibility features of web pages can be helpful in credibility assessment, the amount and type of information should be provided carefully to avoid information clutter for users.

Multiple web page features, including on-page, off-page, and aggregate features were identified by [82] to augment web search results through visualization. The selected features were usually difficult or even impossible for users to extract or assess on their own. This augmentation helped users make more accurate credibility assessment when viewing search results returned by search engines. They used the popularity of the website among experts and general users, awards and certificates won by the website, the number of locations people reach the site from, PageRank, and the domain type of the website as the most promising credibility features. They also collected a dataset of 1000 web pages in selected topics including health, finance, politics, celebrity news, and environmental science and assigned each web page a credibility score based on a five-point Likert scale.

Similarly, [102] proposed a system that calculates and visualizes credibility scores from different credibility aspects such as accuracy, authority, objectivity, coverage, and currency for web search results. Then, the system uses users' feedback to predict their credibility assessment model and re-rank the search results based on the predicted model. Since credibility is a subjective quality, using this method, the system optimized the model for each user adaptively. One other approach to enhance credibility transparency of search results is a system created by [100] that provides users with disputed topics related to their search query as a credibility warning. They investigated how this type of support affects users' search behavior and decision-making regarding the credibility of a web page by measuring clickthrough, dwell time, and page view.

WISDOM is another web credibility analysis system developed by [7] that evaluates the credibility of online information from different perspectives. They analyzed the credibility of information content, information sender, and information appearance. WISDOM investigates the result pages of a search engine for a given query. It presents various automatically extracted information such as a list of major statements and their contradictions about a topic, and a chart of positive and negative opinions distribution. It also displays predefined categories of information authors or publishers such as government, company, academic society, or individuals. However, the provided information might be confusing or overwhelming, making searches longer or more difficult.

2.2.4 Link-based Credibility Ranking Systems

Using web pages link structure to determine a credibility ranking has been a topic of several studies. Ranking algorithms such as PageRank [76] and HITS [54] have been widely used in the information retrieval field to rank web pages based on their popularity and link structure. PageRank determines how relevant and important a web page is to a query considering the number of incoming and outgoing links on the web page that results in a single overall authority score for the page [76]. While these algorithms indirectly indicate the quality of web pages, some researchers worked on similar semi-automatic methods to find credible web pages with a focus on distinguishing good web pages from web spams. Spam pages are usually created to mislead search engines. One common spamming method is to create numerous fake web pages that are pointing to a target web page, and consequently, the number of incoming links of the target web page is increased in ranking algorithms used by search engines [42]. TrustRank system is built based on the presumption that high-quality pages point to other reputable web pages [42]. Therefore, at first, experts identify a small set of high-quality web pages as a seed for the system. Then, the TrustRank algorithm tries to find other web pages that are highly probable to be credible based on the link structure of pre-trusted web pages. CredibleRank is another algorithm proposed by [13] that distinguishes between page quality and link credibility. They argue that it may cause problems if the algorithm is highly dependent on a predefined list of good web pages. They conclude that web page credibility is associated with its distance to a blacklist of spam web pages instead of a whitelist used in [42]. Using this logic, they considered the quality of outgoing links of the web page itself and its neighbor pages that are a few hops away to assess the credibility of a page. In this way, low-quality out-links to spam web pages would decrease the credibility rank of a web page.

2.2.5 General Credibility Assessment Tools

Some researchers have proposed distinct methods to analyze the credibility of various types of online information separately. For example, [91] suggested different methods to assess the credibility of weblogs, images, digital maps and videos based on their content, social support, and author. They introduced particular measures for each type of information on the web. For instance, they analyzed relatedness, consistency, and typicality to assess if an image is suitable for a Wikipedia article. In the case of weblogs, they considered the expertise and sentiment features of bloggers to estimate the credibility of the content. They also used two other factors to estimate the credibility of the web content similar to factors used by [72]: topic coverage and topic depth. Here, topic depth indicates how in-depth the content is from a technical viewpoint. Besides, they analyzed the link structure of the web page to approximate the social support aspect of credibility. They used the majority or dominance metric to indicate how dominant the content is compared to other relevant content. Hence, similarity calculation methods like cosine similarity were used to find how many similar web pages exist for a target web page which indicates the dominance of that web page [91].

WebCAST is another automated web credibility assessment tool developed by [4] that evaluates the credibility of online information based on various factors like popularity, sentiment, update date, and users' previous ratings. They used Multi-Criteria Decision Analysis (MCDA) to find a proper weight for each credibility factor. Finally, each web page among a set of web pages is assigned a relative credibility score by WebCAST.

2.2.6 Machine Learning Approach

The research presented above is key to understand how we can define credibility. Studies that have used machine learning approaches to assess credibility align more closely to the approach taken with the research presented in this thesis. Some studies applied machine learning models to web pages from a certain domain. For instance, [87] used an SVM classifier to directly evaluate the credibility of web pages in the medical domain based on the guidelines provided by Health on Net Foundation known as HONcode principles. Organizations like HON and Quackwatch assess the reliability of websites by manually investigating them to check whether they meet some predefined criteria or not [87]. HONcode principles include guidelines regarding authority, complementarity, privacy, attribution, justifiability, transparency, financial disclosure, and advertising policy. They provide certificates to websites that comply with the eight mentioned principles. [75] is a study that used machine learning techniques to evaluate web page credibility automatically based on two major categories of features, namely content and social features. Content features are extracted either from the textual content of the web page or from web page structure or metadata if available. In contrast, social features demonstrate how popular the web page is and provide information about the link structure of the web page. They applied both binary classification model and regression model with a five-point Likert scale output to assess the credibility of web pages. They obtained a 70% precision and recall score for the Microsoft credibility dataset [82] in binary classification settings.

[96] focused on social and linguistic features extracted from the textual content of web pages to build a machine learning model to predict web credibility and achieved weighted precision between 66% to 70%. In their first experiment, they used General Inquirer, which

is a content analysis tool with a dictionary-backed lemmatizer, to obtain vectors related to psychosocial and psycholinguistic features for each web page in the dataset. General Inquirer dictionary maps word senses to 83 psycholinguistic categories. For the second experiment, they applied a bag of words approach to create the feature vector. Then they applied a logistic regression classification algorithm to predict the credibility of the web pages using obtained feature vectors.

[22] introduced the concept of the bag of tags, which represents the occurrences of HTML tags on each web page. They used HTML2Seq feature as well as other features, such as text category, sentiment from Vader Lexicon, General Inquirer vector, frequency of social tags, domain type, PageRank information, number of outbound links, spam classification, and open-source classification of the web page to achieve the best performance in 2-class and 5-class settings for the C3 and Microsoft dataset.

2.3 Summary

Perceived credibility of online information is a complex, multifaceted concept. There has been a fair amount of research into understanding what factors influence users' perception of website credibility, including work to provide users with an estimate of credibility that they can consider when viewing online information. While there has been some headway made into the autonomous assessment of the credibility of web pages, most of the previous research were either limited to a specific topic or online information source or did not consider features from different credibility aspects. Additionally, the machine learning models that were developed were not validated by human annotators. Since credibility is a perceived quality, creating a model without considering how the model performs on new web pages and validating it with ratings given by actual web users might not reflect the accuracy and generalizability of the model.

Chapter 3

Extracting Credibility Features

3.1 Credibility Categories

3.1.1 Content Quality

The quality of the content has been shown to be a significant indicator of the credibility of a web page [87]. This aspect of credibility usually refers to the writing quality, readability, and clarity of the text content of the web page. The topic coverage and complexity of the content are other factors that affect content quality [84]. Although quality is somewhat a subjective concept, some measures can be used to evaluate it. The following features have been used as factors that are relevant to content quality:

3.1.1.1 Content Length

Content length is shown to be useful in assessing online content quality [58, 49]. Number of words, number of total characters, and number of unique words in the text were calculated to measure this factor.

3.1.1.2 Grammatical quality of the content

The number of question marks and exclamation marks in the content were extracted. This approach is known to provide a rough estimate of the grammatical quality of the content and sometimes can indicate the tone of the writing. [12, 75, 70]

3.1.1.3 Readability

Readability shows the comprehensibility level of the textual content of the page and indicates how difficult it is to read and understand [75, 5]. There are several validated scoring techniques available to analyze text readability. Most of these techniques rely on analyzing the complexity of the vocabulary and the syntax used in the content and statistical features of the text, such as the number of words, sentences, and syllables. While the results of these methods might be covariant, it is common to use different methods at the same time to cover a broader range of content types with different length and various topics because there are subtle discrepancies in the results based on the context of the web page [10, 95]. Online content with middle-range readability scores is more desirable as it can be understood by most of the web users. Five readability assessment methods were used in this research: 1) Flesch Reading Ease Score, 2) Flesch-Kincaid grade level, 3) Simple Measure of Gobbledygook, 4) Dale-Chall readability formula, and 5) FORCAST formula.

1. **Flesch Reading Ease Score (FRES)** rates the difficulty of a text to be read and understood based on a 100-point scale. Higher scores mean that the passage is easier to understand. Passages with scores in the range of 60 to 70 are considered as plain English. They are easily comprehensible by 13 to 15 years old students, while scores below 50 are more difficult to read and need college or university level education. Equation 3.1 shows the formula for this readability score [25].

$$score = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right) \tag{3.1}$$

2. **Flesch-Kincaid grade level (F-K)** is a readability test indicates the number of years of education or the U.S. grade level required to comprehend the text by the user. It correlates almost inversely with FRES. Equation 3.2 indicates how the F-K grade level is calculated [53].

$$grade = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllables}{total\ words}\right) - 15.59 \tag{3.2}$$

3. **Simple Measure of Gobbledygook (SMOG)** estimates the years of education required by the user to be able to understand a piece of writing. SMOG is a gold standard to assess the readability of the health-related content [23] and is calculated by the formula given in equation 3.3 [63].

$$grade = 1.0430\sqrt{number\ of\ polysyllables \times \frac{30}{number\ of\ sentences}} + 3.1291 \tag{3.3}$$

4. **Dale-Chall readability formula** estimates the comprehension difficulty of a text based on the number of difficult words in the content. These difficult words refer to any word other than the list of 3000 words that are easily understood by American students in grade four [15]. The following formula calculates this score:

$$score = 0.1579\left(\frac{difficult\,words}{words} \times 100\right) + 0.0496\left(\frac{words}{sentences}\right) \tag{3.4}$$

5. **FORCAST formula** is a readability score that uses only a vocabulary element. It is especially useful to analyze longer texts with incomplete or fragmented sentences. FORCAST formula is calculated by Equation 3.5 [14]:

$$\text{Grade level} = 20 - (N/10) \tag{3.5}$$

Where N = number of single-syllable words in a 150-word sample.

3.1.2 Authority

Authority is one of the source characteristics that impacts credibility perception of a web page. It refers to the apparent expertise and popularity of the author or publisher of the content and its referential importance [84, 65]. Any information about the author, including the presence of the author's name, credentials, or contact information, can help evaluate the authority of a web page [84]. The author's identity disclosure can be considered as an indicator of trust [80]. It has also been shown to improve the real-world feel of the web page [26].

3.1.2.1 Recognition of Author

If the authors of the content are known, users may identify who put together the content and can contact them for any questions or clarification. Therefore, users tend to perceive it as more credible compared to the content written by an anonymous author [4, 18]. It is especially true for medical websites as users rely on the content published by reputable authors more than others [81]. In addition, the information posted by verified credible users in social media is perceived more credible by the readers [41, 12].

To assess the authority of the web page, the author or publisher of the web page was extracted using the IBM Natural Language Understanding (NLU) author extraction request [2]. Then, a Boolean feature named 'has-author' was defined to check the presence of a known entity as the author of the page as a credibility factor. It returns True if the web page has a known author.

3.1.2.2 Referential Importance

Credible websites are likely to have high referential importance, which means that they have a large number of internal links as other web pages link to them as references [87, 102]. In this research, three backlink metrics were extracted for each web page using the SEOquake service [83]. These metrics include SEMrush backlinks, SEMrush subdomain backlinks, and SEMrush root domain backlinks. They provide the number of links that SEMrush found leading to a web page, its specific subdomain, and the domain as a whole respectively.

3.1.3 Professionalism

The professionalism credibility category deals with perceived professionalism or quality of website design, which is found to be highly influential on perception of credibility [64]. Multiple factors, including the domain type, URL address, presence of advertisements, title, spelling errors, broken links, and visual design of the page, can be considered in evaluating the professionalism of a web page [29, 84]. In this research, the following factors have been used to estimate the professionalism aspect of the credibility: title length and URL length in characters, topic category, and domain type.

3.1.3.1 Topic Category

Topic category indicates the particular category that the web page belongs to based on the frequency of the terms and topics discussed in the content [77]. All the web pages in a specific category have an attribute or a feature in common [96]. Some categories are perceived as more credible than others by users as their presumption of the credibility of different topics is not similar [96]. Also, the users' expectations in terms of content quality, information accuracy, and style of writing differ for each type of web page [96]. For example, it was found that healthy lifestyle categories have lower credibility scores compared to other categories, such as politics and entertainment. [51]. In general, users consider different aspects of credibility when they evaluate web pages of different categories. For instance, in the study conducted by [29], users mentioned website reputation as a factor that they notice more than other factors when they assess the credibility of e-commerce websites. At the same time, they consider information bias and accuracy more in case of review/opinion websites. Furthermore, some types of websites are more likely to have intentionally deceptive content [73].

For this research, the category of each web page was determined by using IBM NLU service, which returns up to three categories assigned to the web page content based on a five-level taxonomy hierarchy [2]. Only level 1 of the category with the highest confidence score was used in this research. The NLU API classifies the web page content into following level 1 categories: art and entertainment, technology and computing, sports, science, business and industrial, automotive and vehicles, style and fashion, health and fitness, travel, society, finance, food and drink, hobbies and interests, home and garden, shopping, education, religion and spirituality, family and parenting, pets, careers, real estate, news, and law, government and politics. Figure 3.1 shows the distribution of categories for the web pages in the dataset that is used in this study; a description of the creation of this dataset can be found in Section 4.1.

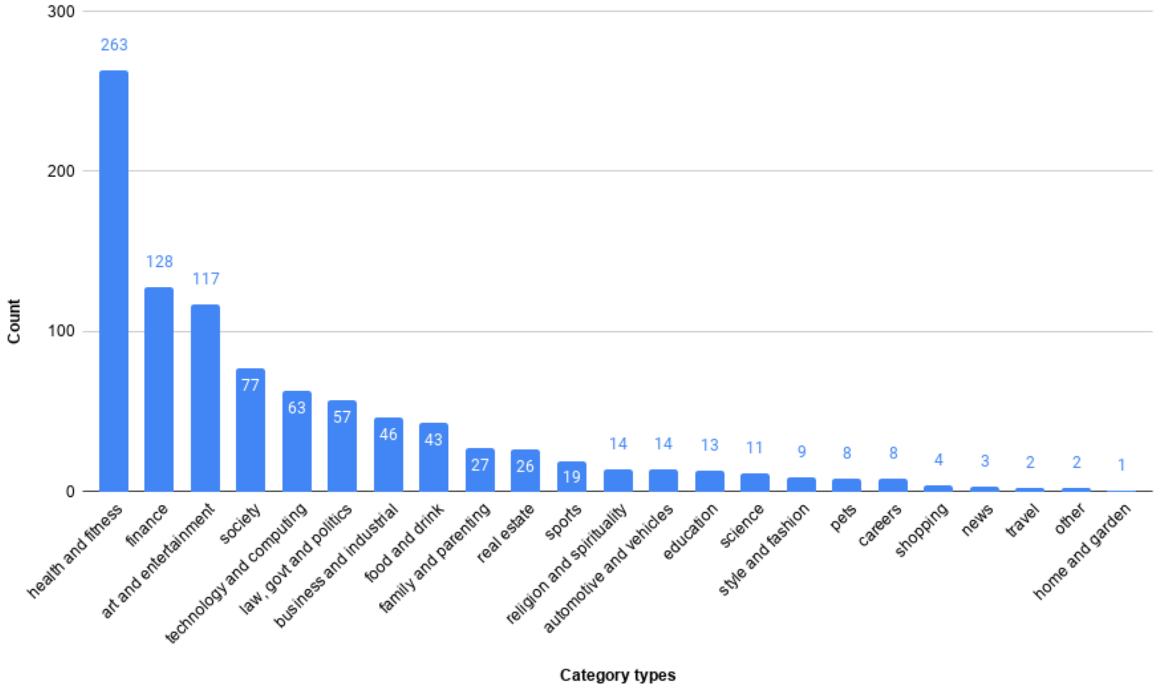

Figure 3.1: Distribution of extracted topic categories in the dataset used in this research.

3.1.3.2 Domain Type

Web page domain type is a factor that may affect the perceived credibility of a web page. A domain suffix is an indicator of the type of website and reflects its purpose to some extent. Usually, people trust websites with academic purposes (e.g., a domain name that ends with an .edu extension) more than websites with commercial purposes [4, 26]. Also, .org or .gov domains that relate to governmental institutions or organizations may contribute to the credibility perception of a web page [75]. The domain type of the web page was extracted from the URL and used as a categorical feature for this research. Figure 3.2 shows the distribution of extracted domain types for the dataset.

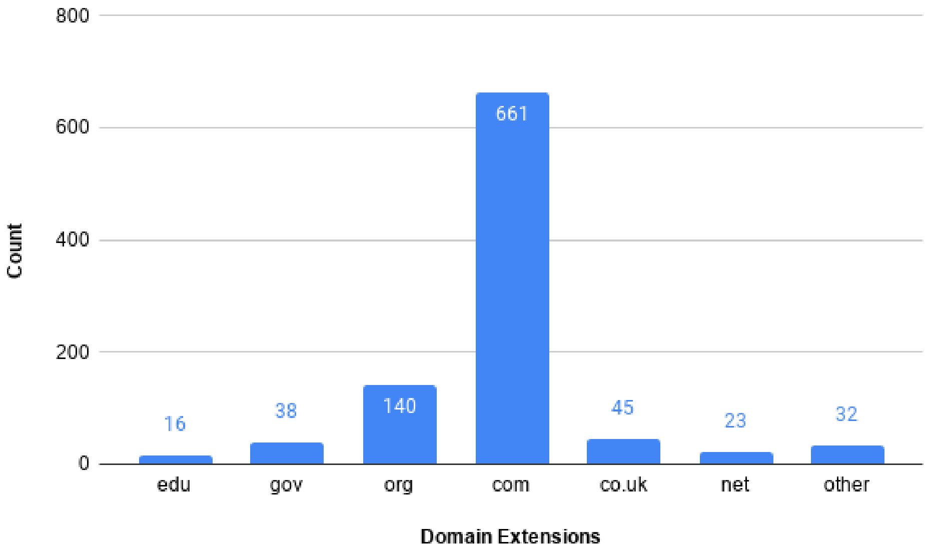

Figure 3.2: Distribution of domain types in the dataset.

3.1.4 Sentiment

Lack of information bias in the web page content, which means that it is written in an impartial and objective manner, can influence the credibility perception of the users [29, 102]. Analyzing the sentiment of the content can reveal whether the content is unbiased (neutral) or it has a positive or negative sentiment [4, 65].

According to [29], unbiased content is perceived as more credible by users. Also, [12] indicated that tweets that contain sentiment terms, particularly positive sentiment, are more likely to have non-credible information. This suggests that there is a relation between perceived credibility and sentiment of the content, therefore, using sentiment features could be helpful in predicting the credibility of a web page.

The document-level sentiment label and sentiment score of the web page were extracted using the IBM NLU API [2]. Sentiment label indicates the polarity of the sentiment, which can be positive, negative, or neutral, and sentiment score shows the strength of the sentiment, which is a value between -1 (negative) to 1 (positive). Also, the list of the web page keywords was extracted. Then, the sentiment associated with the keyword with the highest relevancy score was used as another feature to estimate the impartiality of the web page content.

Another sentiment-based feature is the emotional tone of the content, which may have an impact on the overall credibility perception of the users. The tone of the content can evoke an emotional response from users that might influence their interpretation of the information presented on the web page, particularly in critical domains, such as medical and health-related content [10]. The emotion scores of the content were extracted for five emotions: anger, disgust, fear, joy, and sadness through IBM NLU API [2]. Each score is a value from 0 to 1, with a higher value indicating that the respective emotion is more intense in the text.

3.1.5 Popularity

The popularity aspect reveals how well-reputed a website is among web users [84]. Website popularity and users' past experience with a website are effective factors in how credible the web page is perceived by the users [37, 96]. It is observed that popularity and credibility perception are positively correlated since users assume that web pages that are recommended or viewed by more people are more reliable and authoritative [75, 102, 96]. In addition, the ranking of the web page in a search engine is a good indicator of its popularity [4, 82]. Previously, Google PageRank was used in many studies as an indicator of the relative importance and popularity of a web page, but it is no longer available to the public/developers [87, 22].

The following factors were obtained from SEOquake [83] to measure the popularity category in this study:

1. Alexa rank, which indicates the general popularity of a website. It is the ranking of the website based on its web traffic and page views data.

2. The number of Facebook likes received by the web page that presents its social popularity.

3. The Search engine index was extracted for each web page from multiple search engines, specifically Yahoo, Bing, and Baidu. Google was not used because many web pages returned an error when it was attempted to extract the index for multiple websites; therefore, Google could not be included in this research. This index indicates the number of indexed pages examined by the respective search engine for a given domain. A web page would not be included in search results of a search engine unless that search engine indexes it. To reduce the bias of a specific search engine, indexes from multiple search engines (i.e., Yahoo, Bing, and Baidu) were extracted.

3.1.6 Currency

Currency indicates how up to date the content of the web page is. It can affect the users' credibility perception of the web page as more recent or more frequently updated content seems more credible to them especially in the case of news or time-dependent information [73, 84, 4]. Two factors were considered in this research to evaluate the currency credibility category. First, the publication date of the page was extracted using IBM NLU API [2]. A Boolean feature was defined to check if the web page has a date stamp. If this was the case, the number of years passed since the publication date of the page was defined as a feature. Second, the years passed since the page was first archived by Wayback Machine of Internet Archive was obtained and used as another feature.

3.2 Bag-of-words Features

In addition to the credibility features that were extracted for each web page in the dataset, the term frequency-inverse document frequency vectorizer (TF-IDF vectorizer) from scikit-learn toolkit [1] was used to transform the textual content of the web page to the numeric format required for machine learning models. TF-IDF vectorizer is a popular bag-of-words feature extraction approach that shows how important a word is in a collection of documents. TF-IDF is a method that gives a high weight to any term that frequently appears in a specific document, but not in many documents in the corpus. The TFIDF score for word w in document d is calculated by Equation 3.6:

$$tfidf(w,d) = tf \times \log\left(\frac{N+1}{N_w+1}\right) + 1 \tag{3.6}$$

Where N is the total number of documents in the corpus, N_w is the number of documents with word w, and tf is the count of word w in document d. An important parameter that should be noted when using the TF-IDF vectorizer is the n-gram range. N-gram indicates the number of consecutive words that are considered during vectorizing the text. N-gram range between 1 and 3 was used in this study to capture features for unigrams, bigrams, and trigrams in the content.

3.3 Summary

Six credibility categories and the methods to automatically extract their relevant credibility features were discussed in this chapter. To the best of the author's knowledge, this is the first time several of the extracted features (i.e., FORCAST score, search engine index, URL length, and emotional tone) have been used in a machine learning approach to estimate the credibility perception of web pages.

The credibility categories and related credibility factors that were explored, as well as extracted credibility features for each web page, are summarized in Table 3.1.

Table 3.1: Summary of credibility categories, credibility factors, and credibility features extracted for each web page that were used in this thesis research

Categories	Credibility Factors	Credibility Features
Content Quality	Content Length	Character count, Word count, Unique word count
	Grammatical quality	Question mark count, Exclamation mark count
	Readability	FRES, F-K grade, SMOG, Dale-Chall, FORCAST
Authority	Author recognition	Has-author
	Referential importance	Page, subdomain, and root domain backlinks
Professionalism	Metadata	Title length, URL length
	Topic category	Health and fitness, Finance, Art and entertainment, Society, Technology and computing
	Domain type	edu, gov, org, com, net, other
Sentiment	Document level se	Sentiment label, sentiment score
	Word level	Keyword sentiment label, Keyword sentiment score
	Emotional tone	Anger, disgust, fear, joy, and sadness scores
Popularity	General popularity	Alexa rank
	Social popularity	Facebook likes count
	Search engine index	Yahoo index, Bing index, Baidu index
Currency	Update time	Publication date presence, Web page age
	Update frequency	Internet archive age

Chapter 4

Model Development

4.1 Dataset

The Content Credibility Corpus (C3) [51] was used as a dataset to develop, train, and test the performance of the proposed model in this research. To the author's knowledge, C3 is the largest web credibility dataset that is publicly available for research [51].The C3 dataset was collected over three years and it contains 5543 web pages belonging to five main categories: medicine, healthy lifestyle, politics and economy, personal finance, and entertainment. For approximately 1500 web pages, participants, who were recruited using the Amazon Mechanical Turk platform, were asked to rate the credibility of the websites using a 5-point Likert scale, where a score of five means that the web page is highly credible and a score one is very non-credible. For more details on the C3 dataset, the reader is referred to [51].

Each web page in the C3 dataset was evaluated by multiple participants in the work done by [51]. These evaluations were aggregated into a score that was provided by the dataset; this was used as the credibility score of a web page in this research. Then, each URL address was scraped to retrieve its related web page content. URLs that no longer pointed to a valid web page (e.g., due to a not found error) were removed from the dataset. Finally, a subset dataset of the C3 dataset with 955 URL addresses, along with their respective credibility scores, was obtained. Figure 4.1 shows the distribution of credibility scores in the new subset dataset. It reveals that the dataset is highly imbalanced in favor of web pages with credibility scores of 4 and 5. While this is not ideal, as (to the author's knowledge) no other appropriate datasets are avialable, the subset of the C3 was used to develop the model described in this thesis research.

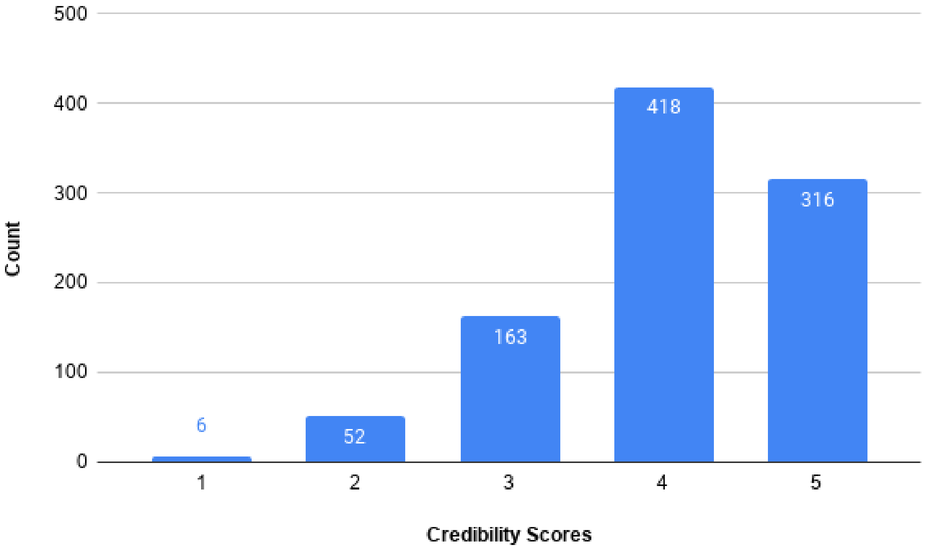

Figure 4.1: Distribution of credibility scores in the subset dataset of the C3 dataset that was used for training the ML models in this research.

4.2 Machine Learning Models

As there are several machine learning (ML) approaches that might be used, this research explored their applicability to the estimation of web page credibility. A brief explication of the different ML models that were considered follows.

4.2.1 Linear Regression

Linear regression is the simplest linear method to predict a value as a weighted sum of the input features. It indicates the relationship between a target value as a dependent variable and features as independent variables. Linear regression tends to optimize by minimizing the mean squared error between the predicted values by the model and the observed target values in the training dataset to establish weight coefficients [31].

4.2.2 Ridge Regression

Ridge regression is another linear model for regression. The formula for predicting the target value is the same as the linear regression. The difference is that Ridge regression uses the L2 regularization (i.e., it penalizes the sum of square weights) to make the coefficients as close to zero as possible to avoid overfitting. A parameter, alpha, is used to adjust the amount of regularization. It controls the trade-off between the performance and simplicity of the model. Increasing alpha leads to better generalization but might reduce the performance of the model on the training set and vice-versa.

4.2.3 Logistic Regression

Logistic regression is a classification algorithm that is used when the target variable is a categorical instead of continuous value. Logistic regression transforms the predicted values to a probability using a sigmoid function. Then the output of the function can be mapped to the different classes [46].

4.2.4 SVM

Support Vector Machine (SVM) is a supervised learning algorithm that is used for classification and regression problems. In the SVM model, each data is represented as a point in n-dimensional space where n indicates the number of features. Then, the algorithm separates the data into different classes by finding a hyper-plane that discriminates classes. The best hyper-plane is considered to be the one that maximizes the margin between classes, which is the distance between the hyper-plane and the nearest data point from each class. The most difficult data points to classify are those that are closest to the decision boundary between classes, which are called the support vectors [19].

An important parameter to tune when using SVM models is the penalization factor (C) that controls the trade-off between the complexity and the classification accuracy of the model. Smaller values of C result in models with lower complexity (i.e., larger margins between hyper-planes) but a greater chance that a data point could be misclassified. Also, Since SVM is not a scale-invariant algorithm, it is necessary to scale the data before using the SVM classifier.

4.2.5 Random Forest

Random forests are ensemble learning methods that consist of multiple moderately different decision trees that are built in order to control the problematic behavior of decision trees that tend to overfit the training data [60]. Each individual decision tree in the random forest might perform a good job in predicting the target, but it may overfit on the part of the data. In random forests, the amount of overfitting is reduced by averaging or soft voting the output of each tree to get the final prediction. The difference between trees in the random forest collection is guaranteed by selecting a random subset of features and bootstrap sampling of the data. The number of trees or estimators to build the random forest can be decided as a parameter in the scikit-learn toolkit [1].

4.2.6 XGBoost

The Extreme Gradient Boosting (XGBoost) algorithm is an implementation of gradient boosted decision trees [33] that has been optimized to have high efficiency of computation time and model performance [17]. It is also a Sparse Aware algorithm that handles missing data values automatically. It can be used for both classification and regression problems.

The XGBoost algorithm is based on decision tree ensemble models. The ensemble model is composed of multiple classification and regression trees (CART) that combines their prediction results. Boosting is an ensemble technique that uses an additive strategy to learn the tree structure. New models are added to correct the prior model error until no further improvement is possible. Gradient boosting uses the gradient descent algorithm to minimize the differentiable loss function. Also, Gradient boosted trees usually use trees with low depth as weak learners. Strong pre-pruning is used in gradient boosting trees instead of the randomization approach used in the random forest model. Similar to the random forest, the complexity of the model is increased by adding the number of estimators. Gradient Boosted trees work well on a combination of continuous and categorical features.

4.3 Oversampling Methods

Since the subset of the C3 dataset that was used in this research is imbalanced, the performance of the classifiers to predict the minority class (i.e., the low credibility class) is lower compared to the majority class. Two oversampling techniques were used to address the imbalanced dataset problem:

1. Synthetic Minority Oversampling Technique (SMOTE): SMOTE augments the data by synthesizing new samples from existing samples. It uses the k-nearest-neighbor method to find the nearest samples to a random sample of the minority class. Then, it creates a synthetic sample at a randomly selected point between two examples of the minority class in the feature space. The synthetic samples are similar to the minority class samples, but they are slightly modified [16].

2. Adaptive Synthetic Sampling (ADASYN): ADASYN is a modified version of SMOTE that considers the distribution of minority class samples to generate the synthetic samples. It adaptively creates more synthetic samples of the minority class samples that are harder to learn. It means more samples are generated in the regions of feature space where there are few samples of the minority class compared to the regions that the density of minority class samples is higher [44].

4.4 Evaluation Metrics

The classification performance of the models was evaluated using accuracy, precision, recall, and F1 score metrics. Since the subset of the C3 dataset used in this study is imbalanced, the weighted average for each metric was obtained as well.

1. **Accuracy** is the number of correct predictions divided by the total number of predictions. It is calculated by the formula in Equation 4.1, where FP is the number of incorrect positive predictions or false positives. The number of incorrect negative predictions or false negatives is denoted FN. TP and TN correspond to the number of correct positive and negative predictions, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

2. **Precision** is the measure of exactness of the model and indicates how many of the samples predicted as positive are actually positive and is calculated by Equation 4.2.

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

3. **Recall** is the number of positive predictions divided by the number of positive samples in the dataset. Equation 4.3 shows the formula to calculate recall.

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

4. **F1 score** is a harmonic mean of precision and recall and is calculated by Equation 4.4.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.4}$$

The following metrics were measured to evaluate the performance of regression models:
1. **Coefficient of determination ($\mathbf{R^2}$)**, which provides an indication of the goodness of fit of the regression model and is calculated by Equation 4.5. $\hat{y}_i$ is the predicted value for i-th sample, y_i is the corresponding observed value for that sample, and n is the total number of samples.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.5}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

2. **Root Means Squared Error (RMSE)** calculates the differences between the predicted values by the regression model and observed values. It corresponds to square root of the average of squared errors. Equation 4.6 shows the formula for RMSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.6}$$

3. **Mean Absolute Error (MAE)** that corresponds to the expected value of the absolute error loss or $l1$-norm loss. MAE is calculated by Equation 4.7.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.7}$$

4.5 Feature Selection

First, features from different credibility categories were extracted for each web page in the dataset according to the methods explained in Chapter 3. Then, one-hot encoding was used to replace the categorical features, such as category type, domain type, and sentiment labels with new features with values 0 and 1 as per category. Overall, 75 features were compiled after using the one-hot encoder. To reduce the dimensionality of the feature set, a one-way ANOVA feature selection was applied to select the most significant features for the classification task. The F-value feature selection function was used as a univariate linear regression test to select the top features for the regression experiments using scikit-learn toolkit [1]. Experiments were performed using 10, 20, 40, 60, 80, and 100 percentiles of the features selected by one-way ANOVA or F-value to select the optimized subset of features for each model.

Three sets of features were used in each classification experiment to compare how three features sets impact the model performance:

1. Extracted credibility features
2. TF-IDF feature vector
3. Combination of extracted features and TF-IDF features

Since classifiers such as SVM are sensitive to the scaling of the data, the standard scaler in the scikit-learn toolkit was used to bring all the features to almost a similar range of magnitude. Standard scaler transforms the feature in a way that makes sure it has the mean equal to 0 and the variance equal to 1 [1].

4.6 Machine Learning Experiments

Two main settings were used to autonomously estimate the credibility of web pages in this study:

1. **Classification** was treated as a binary variable where web pages were deemed to be credible or non-credible. Web pages with credibility scores of 4 and 5 were labeled as credible ("High") and web pages that were rated lower or equal to 3 were labeled as non-credible ("Low"), as has been done in previous studies [75, 96, 22]. This approach was chosen because it has been shown that it is difficult to define a third or more separate class(es) for web pages with medium credibility as the criteria for such class is not clear [87]; in other words, people are more agreed on what is 'credible' or 'not credible', but have high variance with 'moderately credible'.

2. **Regression** was used to estimate the level of credibility of a web page as a numeric value based on a 5-point Likert scale.

The scikit-learn toolkit [1] was used to build the machine learning models in this study. Several machine learning models, including logistic regression, SVM, random forest, and XGBoost classifier for the classification task, and linear regression, ridge regression, SVM regressor, random forest regressor, and XGBoost regressor for the regression task were tested. In each experiment, a grid search was used to find the optimal parameter combination for the model. In all experiments, the dataset was split into an 80% training set and a 20% test set. Stratified three-fold cross-validation was performed to evaluate the generalization performance of the model.

4.7 Results

4.7.1 Top Features

Table 4.1 shows the top 10 selected features using one-way ANOVA for the classification task and F-value for the regression task. Features from all credibility categories except sentiment category are among the top features selected for the classification task. Features that belong to the professionalism category constitute the majority of the top features.

Table 4.1: Top 10 features selected by ANOVA/F-value feature selection for the Classification and Regression tasks

	Classification Task		**Regression Task**	
Rank	**Feature**	**Category**	**Feature**	**Category**
1	Has Author	Authority	Domain Type (gov)	Professionalism
2	Alexa Rank	Popularity	Category (Careers)	Professionalism
3	SMOG Score	Content Quality	Alex Rank	Popularity
4	FORCAST Score	Content Quality	FORCAST Score	Content Quality
5	Web Archive Age	Currency	SMOG	Content Quality
6	Category (Tech.)	Professionalism	Web Archive Age	Currency
7	Domain Type (gov)	Professionalism	Has Author	Authority
8	Domain Type (net)	Professionalism	Domain Type (net)	Professionalism
9	URL Length	Professionalism	Sentiment Label (Neg.)	Sentiment
10	Bing Index	Popularity	Sentiment Score	Sentiment

The top 30 coefficients with the largest absolute value of the logistic regression model that was trained on TF-IDF features and their corresponding unigram and bigram terms were extracted and are shown in Figure 4.2. Positive coefficients indicate the terms that are most prevalent for credible web pages according to the model. In contrast, negative coefficients are associated with terms that belong to non-credible web pages.

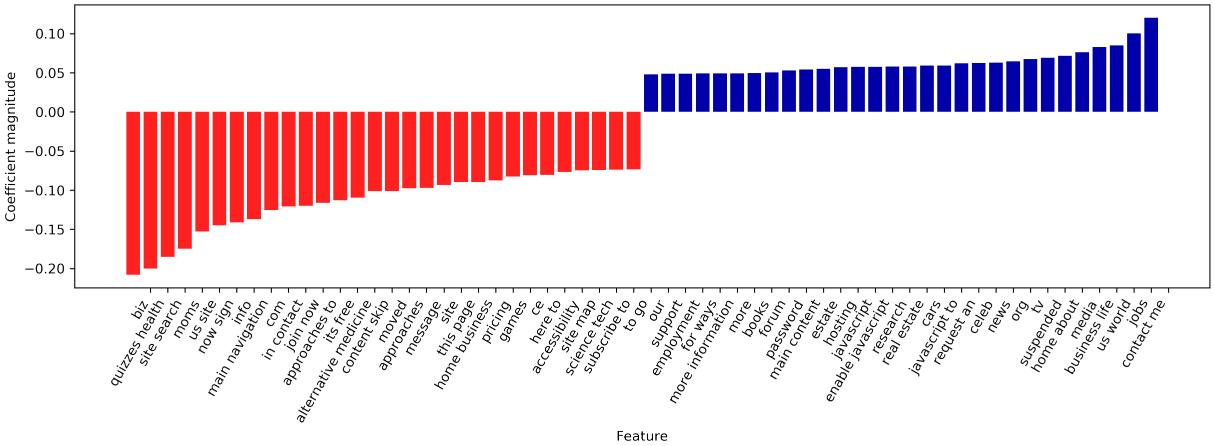

Figure 4.2: Unigram and bigram terms with the 30 largest coefficients in a logistic regression model.

4.7.2 Classification Results

Figure 4.3 shows the weighted average F1 score for all classifiers trained on three different feature sets: 1) extracted features; 2) TF-IDF features; and 3) all features. Overall, the results obtained for the classifiers using different ML methods were slightly different. The best performance based on the weighted average F1 score was obtained with XGBoost for extracted features. In comparison, the Logistic regression classifier had the best performance when TF-IDF features and all features together were used as a feature vector.

4.7.2.1 Logistic Regression

Table 4.2 shows the result of the experiment using logistic regression classifier. Percentile and number of the features, as well as the best parameter combination that led to the highest performance, were determined by grid search and are reported in the first column. The results of the model trained on TF-IDF features and all features were similar and

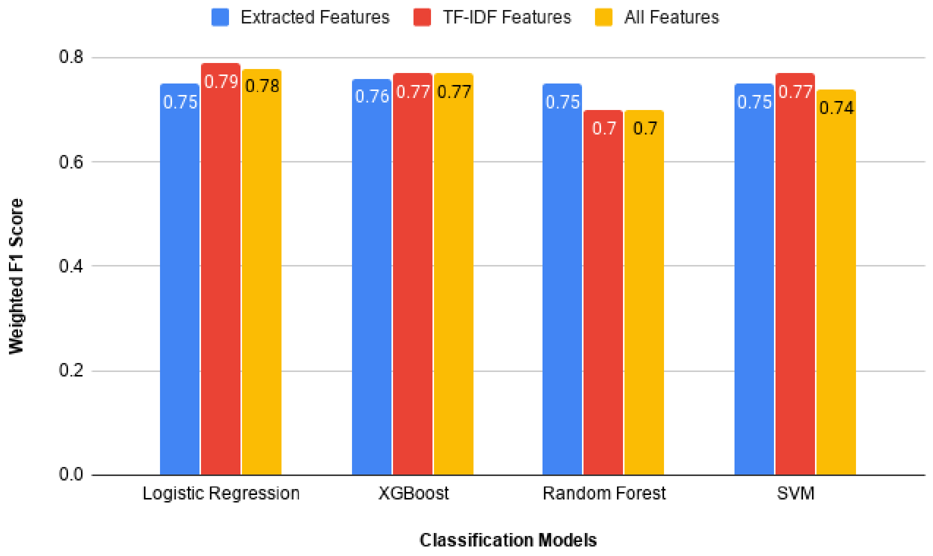

Figure 4.3: Weighted average F1 score for all classifiers trained on three feature sets.

outperformed extracted features, but the number of required features for the best performance was much less with all features compared to TF-IDF features only. This indicates that some of the inferred content characteristics by TF-IDF features have already been captured by the extracted credibility features.

Table 4.2: Logistic regression classification results.

Features	Class	Precision	Recall	F1 Score
Extracted (100%, n = 75) C = 1	**Low**	0.43	0.29	0.35
	High	0.82	0.89	0.86
	Weighted	0.74	0.76	0.75
	Accuracy	0.76		
TF-IDF (80%, n = 10623) C = 0.1	**Low**	0.61	0.34	0.44
	High	0.84	0.94	0.89
	Weighted	0.79	0.81	0.79
	Accuracy	0.81		
All Features (20%, n = 2670) C = 0.1	**Low**	0.54	0.37	0.43
	High	0.84	0.91	0.88
	Weighted	0.78	0.88	0.78
	Accuracy	0.80		

4.7.2.2 XGBoost

The results obtained for the XGBoost classifier are shown in Table 4.3; the best combination of the number of estimators and the max depth of the tree is reported in the first column. The weighted F1 score for the model trained on extracted features is the highest among all classifiers with the lowest percentile of features. The F1 score achieved for low credibility class is higher in the XGBoost classifier compared to other classifiers, while the model is trained on extracted features.

Table 4.3: XGBoost classification results.

Features	Class	Precision	Recall	F1 Score
Extracted	**Low**	0.45	0.37	0.41
(10%, n = 8)	**High**	0.84	0.88	0.86
n_estimators = 100	**Weighted**	0.75	0.77	0.76
max_depth = 4	**Accuracy**	0.77		
TF-IDF	**Low**	0.73	0.20	0.31
(40%, n = 5312)	**High**	0.82	0.98	0.89
n_estimators = 100	**Weighted**	0.80	0.81	0.77
max_depth = 3	**Accuracy**	0.81		
All Features	**Low**	0.55	0.27	0.36
(60%, n = 8012)	**High**	0.82	0.94	0.88
n_estimators = 100	**Weighted**	0.77	0.80	0.77
max_depth = 3	**Accuracy**	0.80		

4.7.2.3 Random Forest

Table 4.4 summarizes the results for the random forest classifier. As expected, random forest performed poorly when it was trained on sparse features (i.e., TF-IDF features). Max depth and the number of required estimators to achieve the best performance are also higher compared to XGBoost.

Table 4.4: Random forest classification results.

Features	Class	Precision	Recall	F1 Score
Extracted	**Low**	0.47	0.22	0.30
(60%, n = 45)	**High**	0.81	0.93	0.87
n_estimators = 200	**Weighted**	0.74	0.78	0.75
max_depth = 7	**Accuracy**	0.78		
TF-IDF	**Low**	0.50	0.02	0.05
(40%, n = 5312)	**High**	0.79	0.99	0.88
n_estimators = 200	**Weighted**	0.73	0.79	0.70
max_depth = 7	**Accuracy**	0.79		
All Features	**Low**	0.33	0.02	0.05
(10%, n = 1336)	**High**	0.79	0.99	0.88
n_estimators = 100	**Weighted**	0.69	0.78	0.70
max_depth = 7	**Accuracy**	0.78		

4.7.2.4 SVM

The results for the SVM classifier are reported in Table 4.5. SVM performed well, particularly for the low credibility class, when it was trained on TF-IDF features and combined features.

Table 4.5: SVM classification results.

Features	Class	Precision	Recall	F1 Score
Extracted (100%, n = 75) C = 1	**Low**	0.43	0.29	0.35
	High	0.82	0.89	0.86
	Weighted	0.74	0.76	0.75
	Accuracy	0.76		
TF-IDF (10%, n = 1328) C = 0.001	**Low**	0.49	0.41	0.45
	High	0.85	0.88	0.86
	Weighted	0.77	0.0.78	0.77
	Accuracy	0.78		
All Features (10%, n = 1336) C = 0.0001	**Low**	0.41	0.71	0.52
	High	0.90	0.72	0.80
	Weighted	0.79	0.72	0.74
	Accuracy	0.72		

4.7.2.5 Oversampling Results

Table 4.6 demonstrates the results of experiments that were conducted to improve the performance of the model for the low credibility class. Two classifiers that achieved better performance in the classification task, i.e., logistic regression and XGBoost were chosen as the classifiers. Also, TF-IDF features and all features were used as feature vectors. SMOTE and ADASYN were the two oversampling techniques that were applied to achieve optimal results.

The best performance was obtained with the logistic regression classifier (C = 0.1). Using oversampling improves recall and the F1 score for the low credibility class. On the other hand, the weighted average F1 score and accuracy of the model decreased due to the decrease in recall for the high credibility class. This means that the user must consider what type of misclassification is more critical, which requires knowledge of the application of the model, when deciding whether oversampling should be applied or not.

Table 4.6: Results for the oversampling experiment.

Method/Features	Class	Precision	Recall	F1 Score
SMOTE **TF-IDF** **(20%, n = 2655)**	**Low**	0.44	0.71	0.54
	High	0.90	0.75	0.82
	Weighted	0.80	0.74	0.76
	Accuracy	0.74		
ADASYN **All Features** **(20%, n = 2670)**	**Low**	0.44	0.61	0.51
	High	0.88	0.79	0.83
	Weighted	0.79	0.75	0.76
	Accuracy	0.75		

4.7.3 Regression Results

As using sparse textual features did not lead to any meaningful results for the regression task, only extracted credibility features were used to train the regression models. Error and goodness of fit metrics for different models are compared in Table 4.7. The best parameter combination chosen by grid search and optimal percentile of features are demonstrated in Table 4.7. The highest coefficient of determination and the lowest RMSE value were obtained with the random forest regressor; the SVM model achieved the lowest MAE value.

Table 4.7: Results for the regression experiments.

Regression model	R^2	RMSE	MAE	Parameter	Feature %
Linear Regression	0.1	0.77	0.6	-	10
Ridge Regression	0.1	0.77	0.59	Alpha = 100	10
SVM	0.09	0.77	**0.57**	C = 0.1 Kernel = rbf	10
XGBoost	0.11	0.76	0.6	max_depth = 3 n_estimators = 40	10
Random Forest	**0.14**	**0.75**	0.6	max_depth = 4 n_estimators = 50	20

4.8 Observations and Discussion

- The list of top features in Table 4.1 indicates that both content features and features related to the source characteristics or metadata of the web page are useful in estimating its perceived credibility. Using word features has an advantage in that it eliminates the need to rely on external APIs. On the other hand, word features do not capture off-page features that are unrelated to the textual content of the page, such as Alexa rank, which is one of the significant predictors of the credibility of a web page as indicated in Table 4.1. Thus, although models trained on only word features (TF-IDF features) can perform well, they miss page characteristics such as these. Therefore, using the combination of extracted credibility features and word features can lead to a more robust model that considers various aspects of credibility. Also, extracted credibility features from established credibility categories are more meaningful, easier to interpret, and can contribute to the transparency of the model.

- Two readability scores are among the top selected features: SMOG and FORCAST. It is possibly because of the fact that 'health and fitness' is the prevalent topic category among the topics in the dataset used in this research, and SMOG has shown to be a good measure to evaluate health-related content [23]. FORCAST is useful in assessing the readability of textual content with incomplete sentences. Since web pages usually have different textual elements other than the main content, that might have fragmented sentences, FORCAST can be a good predictor if their readability. Flesch-Kincaid grade level was also among the top 20 features that were selected by one-way ANOVA feature selection. While SMOG and FORCAST were the most promising readability scores in the credibility classification task in this research, other calculated readability scores might be useful in measuring readability depending on the context of the web pages.

- Figure 4.2 shows how word features can imply some of the credibility categories. The 'contact me' bigram, which has the highest coefficient, relates to the authority category. Interestingly, the top selected feature in the classification task is 'has author' that belongs to the authority category as well. Also, the list of top terms indicates that specific topics or domains are more likely to be perceived as credible by users. Words such as 'celeb', 'media', and 'tv' are associated with the web pages about the topics that are perceived as more credible by users. On the other hand, health-related keywords such as 'alternative medicine' turned out to be indicators of web pages that are perceived non-credible. Terms with a commercial connotation such as 'its free', 'join now', 'home business' are among keywords that have a higher coefficient in predicting non-credible web pages. Therefore, the topic category of the web page seems to be a valid feature in predicting the perceived credibility of web pages. However, not all word features can be interpreted easily, and some might not seem that meaningful (e.g., 'ce', 'more', 'our'). Also, using different tokenization or stemming methods or changing the list of stop words might lead to a different set of features and performance. Aspects such as these would need to be explored in more depth prior to implementation with real users.

- The model proposed by [22] achieved the weighted average F1 score of 0.674 for binary classification on the C3 dataset, while the obtained F1 score for the developed model in this research was 0.76 for extracted features, and 0.78 for all features. However, since some of the web pages were removed from the C3 dataset due to unavailability, the accuracy of this comparison might be affected.
- While the proposed model performs well for credible class, the F1 score for non-credible class is lower, which means it is more likely that the model will misclassify non-credible web pages. Although the oversampling technique improved the F1 score for the low credibility class, non-credible web pages are still more prone to misclassification.

4.9 Summary

Several machine learning models were trained on three different feature sets, including automatically extracted credibility features, word features, and combined features. Some of the features explored in this research are being used for the first time, such as FORCAST score, search engine index, and URL length; these were among the top selected features by the feature selection methods. This indicates that these features are good predictors of web page credibility. The best performance was obtained by XGBoost for extracted features and logistic regression for combined features; both outperformed previous work in peer-reviewed literature on the credibility classification task on the C3 corpus. Using oversampling methods improved recall and the F1 score for the low credibility class, which led to less misclassification of non-credible web pages as credible ones. It is important when accurately identifying non-credible web pages is more critical.

The random forest regressor achieved higher R^2 (0.14 vs. 0.133) and lower RMSE (0.75 vs. 0.92) and MAE (0.60 vs. 0.74) compared to the previous work done by [22]. These results suggest that a random forest regressor with extracted features is the best regression approach for autonomous estimation of website credibility for the combinations that were explored in this thesis research.

Chapter 5

Human Validation Experiment

As described in Chapter 4, the model that was selected for this research is a logistic regression classifier trained on all features. To explore the legitimacy of this model, a crowdsourcing task was created and conducted to evaluate to what extent the output of the proposed model aligns with the credibility ratings given by human annotators. This was used to gain an estimate of how accurate the model is in estimating users' perceived credibility.

5.1 Selecting Web Pages

Users' web page credibility perception has a dynamic nature; it can vary over time because of the changes in the information source or user's beliefs. Also, different people can perceive the credibility of a web page differently. Since the web pages in C3 corpus were collected before 2017, it was of interest to find out how current users evaluate the credibility of the web pages and how their ratings correlate to the ratings that were used as ground truth to train the model. Therefore, 30 web pages were selected from the dataset that the algorithm had been trained and tested on; these pages were sampled from different categories, including healthy lifestyle, medicine, personal finance, and entertainment. The distribution of the credibility scores for the selected web pages was kept the same as the original dataset.

In addition to the 30 pages from the C3 dataset, 30 new web pages were selected to explore how generalizable the algorithm is for classifying new web pages. These pages were captured from the websites that were in the top 30 Google search results for the following queries: “Alzheimer’s cure”, “dementia psychic cure”, “extreme workout”, “extreme diet”, “investing in stocks”, and “video game fatigue”.

The queries and web pages were chosen in a way that both credible and non-credible web pages on the same topic could be provided to the annotators. Since the model performed well in predicting credible web pages in the original dataset, a more balanced set of credible and non-credible web pages were selected for the validation experiment. This allowed us to explore how the model performs overall when there is more of a balance of credible to non-credible web pages. Web pages were screened to ensure they did not contain inappropriate content. In total, 60 webpages (30 from the C3 dataset, 30 new ones for this research) were used in the crowdsourcing task.

5.2 Participants

This study was reviewed and received ethics clearance through the University of Waterloo Office of Research Ethics (ORE#41767).

Participants were recruited through the appen crowdsourcing platform [1] (formerly known as Figure Eight). Only adults aged over 18 years old were included as participants because they are the age group that can register as a contributor in the appen platform. They were required to be able to read and write in English because the content of selected web pages and questions were in English.

Appen provides the task to its registered contributors. Contributors can select a project from the task wall if they are interested in that and start doing the task after reading the instructions. There are three levels of contributors in the appen platform based on the quality of their previous works. Level 3 was chosen for this study, which includes the most experienced, highest quality contributors. The minimum time that should take a contributor to evaluate each web page and answer the questions was set to 60 seconds to help ensure they did not randomly answer the questions without investigating a web page. Answers that took less than the minimum time were automatically discarded from the data.

[1] www.appen.com

5.3 Method

First, participants completed the online consent form and answered the demographic questions regarding their age, gender, education level, and the country they were participating from. Then, participants were asked to click on a link to a web page from the list and answer the following questions regarding the credibility of that web page:

1. How credible do you think the web page is based on a five-point scale rating? (1: Very non-credible, 5: Very credible)

2. Please briefly explain why you chose the credibility rating for this web page that you did.

3. How knowledgeable do you consider yourself to be about the topic of the web page? (1: Not knowledgeable at all, 5: Highly knowledgeable)

Participants were randomly provided five web pages to evaluate at each step of the task. They needed to annotate all the web pages in each step to go to the next step. They were allowed to evaluate up to 30 web pages in six steps. Ten judgments were collected for each of the 60 web pages (30 from C3, 30 new for this research).

5.4 Results

5.4.1 Demographics

69 participants (56 male; 13 female), evaluated the credibility of the 60 web pages. Each page was evaluated 10 times for a total of 600 evaluations. Participants had an average age of 27.43 (SD = 8.86); the oldest and youngest participants were 55 and 18 years old, respectively. Most of the participants (84%) were from the United States. Table 5.1 shows the number of participants from each country.

Table 5.1: Participants' location

Country	**U.S**	**Venezuela**	**Egypt**	**Canada**	**Peru**
Number of Participants	58	6	3	1	1

The distribution of the education level of participants is shown in Figure 5.1. Most of the participants reported a college degree as their highest completed level of education.

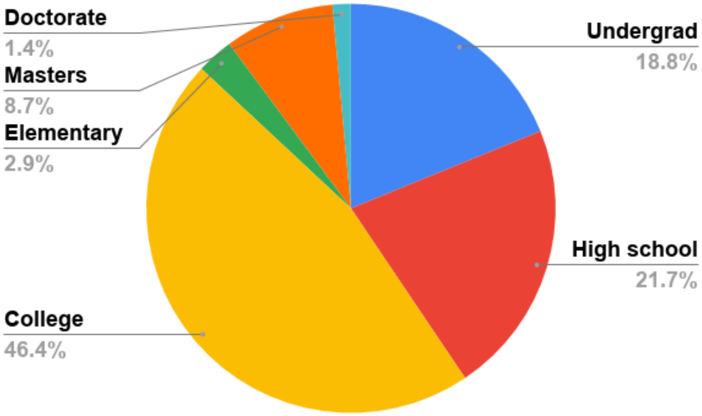

Figure 5.1: Participants' reported highest level of education.

5.4.2 Aggregated Results for All Web Pages

To estimate the inter-rater reliability of the ratings given by participants, Cronbach's alpha reliability coefficient was calculated for all 600 evaluations and was found to have alpha = 0.923. An alpha greater than 0.7 indicates a high agreement between human annotators.

The results for credibility score and knowledgeability score are reported in Table 5.2 for all 600 evaluations. Skewness shows how symmetric the distribution of data is. The skewness value for a normal distribution is 0, while a negative value indicates more weight in the right tail of the distribution. The Shapiro-Wilk (S-W) test was used to check whether the scores were normally distributed. S-W test compares data to a normal distribution with the same mean and standard deviation. If the test is not significant, data is normally distributed. The P-value of S-W test results indicated that the credibility and knowledgeability scores were not normally distributed.

Table 5.2: Credibility and knowledge score for all 600 evaluations

Score	**Count**	**Mean**	**Std**	**Median**	**Skewness**	**Shapiro test**
Credibility Score	600	3.480	1.185	4	-0.235	0.883, p <0.001
Knowledge Score	600	3.26	1.192	3	-0.038	0.895, p <0.001

Since scores were not normally distributed, the correlation between credibility scores and knowledgeability scores was calculated using Spearman's rank correlation coefficient (Spearman's ρ). Spearman's ρ is a non-parametric measure of rank correlation, which evaluates the monotonicity of the relationship between two variables. The Spearman's ρ correlation is 0.52 ($p < 0.0001$), which shows an apparently statistically significant positive correlation between users' knowledgeability of the topic and their associated credibility ratings. This means that users' perceived knowledgeability about a topic can be an effective factor in how they assess the credibility of a web page.

5.4.3 Aggregated Results for Web Pages from C3 dataset

The average and mode of credibility scores given by participants were calculated for each of 30 URLs from the C3 dataset. For ratings with multiple modes, the average of the modes was computed. Average ratings were selected as the final aggregated rating for the URL. Figure 5.2 shows the average of ratings given by participants to each URL along with the original credibility score of the page in the C3 dataset.

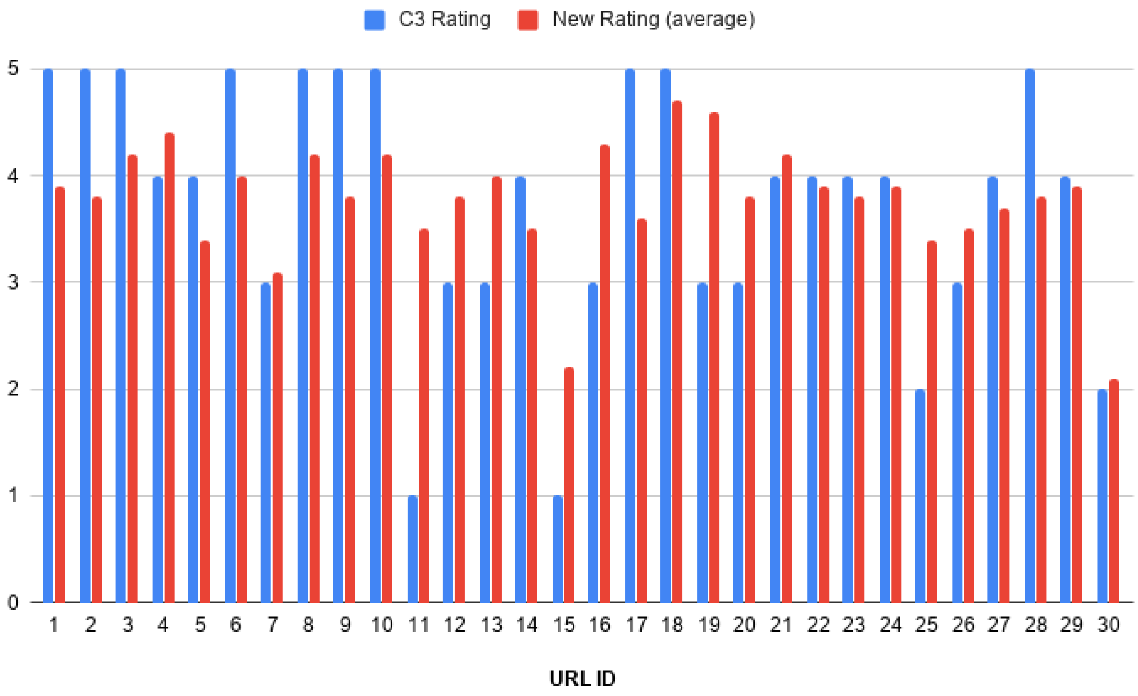

Figure 5.2: Credibility ratings given by participants and credibility scores in C3 dataset for the 30 URLs used in this thesis research.

Spearman's ρ between average credibility scores given by participants and original scores in the C3 dataset was calculated to determine the correlation between these two ratings: $\rho = 0.44, p < 0.02$ indicates a moderate positive correlation between participants' ratings and original scores in the C3 dataset. On average, the original ratings for the C3 dataset were slightly lower than the ones from the new crowdsourcing task. The average of the original ratings for the selected web pages was 3.766 compared to 3.773 for the new ratings given by the crowdsourcing annotators in this research.

5.4.4 Aggregated Results for New Web Pages

The average credibility score for each URL in the list of new web pages was calculated. The median of these averages was 3.2, and the mean was 3.186. Since the average credibility score is a continuous value, to compare the rated results with binary output of the classification model, the web pages with an average rated score above 3.2 were labeled as credible (High) and other web pages were labeled as non-credible (Low). The obtained results are reported in Table 5.3 and the contingency table as shown in Table 5.4 compares the predicted scores by the model with the rated scores by crowdsource annotators.

Accuracy, precision, recall, and F1 score were calculated and are reported in Table 5.5. The number of false positives (FP = 4) was twice the number of false negatives (FN = 2). This means that more non-credible web pages were predicted as credible in comparison to credible web pages that were labeled as non-credible.

5.5 Discussion

Overall, the model performed relatively well on the new web pages outside of the training dataset (Accuracy = 0.80; F1 score = 0.77), which indicated the generalizability of the proposed model. Thus, the autonomous rating of perceived credibility appears to be a promising approach. In particular, the following key points are noted:

- High inter-reliability agreement between users ratings could result from the criteria that was defined for recruitment of participants from the most experienced, highest quality contributors in the appen platform to achieve more accurate ratings.
- Current users ratings and the original ratings in the C3 dataset are moderately correlated, with a slightly different mean. While the subset of pages selected from

Table 5.3: Rated scores and predicted scores for new web pages

URL ID	Avg. Rated Score	Binary Rated Score (human annotator)	Predicted Score (autonomous algorithm)
1	2.7	Low	Low
2	2.6	Low	Low
3	2.7	Low	Low
4	2.6	Low	Low
5	2.9	Low	Low
6	2.8	Low	Low
7	4.2	High	High
8	3.3	High	Low
9	2.9	Low	High
10	3.1	Low	Low
11	3	Low	Low
12	3.1	Low	Low
13	3.2	Low	Low
14	3.5	High	Low
15	3.4	High	High
16	2.9	Low	High
17	3.8	High	High
18	3.2	Low	Low
19	3.7	High	High
20	2.7	Low	High
21	3.3	High	High
22	3.2	Low	Low
23	3.6	High	High
24	2.9	Low	High
25	3.2	Low	Low
26	3.3	High	High
27	3.5	High	High
28	3.7	High	High
29	3	Low	Low
30	3.6	High	High

Table 5.4: Contingency table of predicted scores by the model vs. rated scores by participants.

Predicted **Rated**	**High**	**Low**	**All**
High	10	2	12
Low	4	14	18
All	14	16	30

Table 5.5: Performance metrics of the model for new web pages.

Accuracy	**Precision**	**Recall**	**F1 Score**
0.80	0.71	0.83	0.77

the C3 corpus was from all main categories, a higher number of pages would better show if the previous ratings are still valid considering the changes in users credibility perception and information of web pages over time.

- Average scores for four web pages that were rated as non-credible, but misclassified as credible by the model (i.e., false positives) were: 2.9, 2.9, 2.7, 2.9. On the other hand, the average scores for the 2 false negatives were 3.3, 3.5. Most of these average scores are close to the threshold that was set for classifying credible and non-credible web pages (i.e., 3.2). It means that improvements in the model would probably improve the accuracy of the prediction.

- Ratings given by human annotators were subjective because of the nature of the task. Therefore, misclassifications by the model, especially when it is close to the classification cutoff value, may be due to differences because of subjective perceptions.

5.6 Summary

In this chapter, the crowdsourcing task that was conducted to evaluate the legitimacy and generalizability of the model was described. Sixty web pages, including 30 web pages from the C3 dataset and 30 new web pages, were evaluated by 69 annotators. In total, 600 evaluations were collected. To the author's knowledge, no other study has validated the output of a machine learning-based credibility classification model for entirely new web pages outside of the training dataset. Results indicated a moderate positive correlation between original ratings in the C3 dataset and average ratings given by new annotators. Namely, the proposed model achieved an accuracy of 80% and an F1 score of 0.77 for new web pages, which indicated that the model generalizes relatively well. These results suggest that the proposed model could be helpful in general applications of autonomous estimation of perceived credibility.

Chapter 6

Conclusions, Contributions and Future Work

6.1 Conclusions

While estimating the perceived credibility of web pages is a challenging task due to the combined subjective and objective nature of credibility, this research has put forward a machine learning-based model to automatically classify web pages in terms of perceived credibility.

As different people notice different aspects while evaluating the credibility of a web page, it was deemed necessary to consider various perspectives while building a general-purpose credibility assessment model. Numerous content features, as well as source features that are not directly inferable from the content, were identified and extracted for each web page to cover six different credibility categories, including content quality, authority, professionalism, sentiment, popularity, and currency. Several machine learning experiments were conducted to determine what combination of features and model parameters should be used. The best performance was obtained by using XGBoost classifier for extracted credibility features and logistic regression for text features and combined features. The developed model achieved higher accuracy and F1 score compared to the previous work in credibility assessment using the C3 dataset done by [22]. As the results of this research indicate, features from different credibility categories are influential and required in predicting the perceived credibility of a web page.

A crowdsourcing task was conducted to compare the output of the proposed model with credibility ratings given by human annotators for a set of new web pages. Results achieved an 80% accuracy in estimating users average credibility ratings, which indicates good potential for the model to generalize well. Classifying non-credible web pages correctly was observed to be more difficult than credible ones, however, a larger sample size is required to explore this in more depth.

It is important to note that while automated perceived credibility prediction models are helpful to users, like other automated systems, they are not completely error-free and accurate. Therefore, users need to avoid overtrusting the system (i.e., putting too much trust in the system), and ways of ensuring that users are aware of the limitations of the system are made clear. This can be more critical when non-credible web pages are misclassified as credible, which might lead users to trust misleading information. Thus, the output of the system should always be portrayed to the user in a way that they can carefully and thoughtfully consider what the rating means.

6.2 Contributions

To the author's knowledge, the research presented in this thesis adds the following to the body of knowledge:

1. A comprehensive set of credibility features from different credibility categories were used to build the model. Some of the features were presented in this research for the first time and were observed to be promising in classifying the credible web pages.

2. The classification and regression models developed in this thesis achieved higher performance and lower error scores compared to the previous work that has been done using the C3 dataset.

3. A set of 30 new web pages from different topic categories were selected to validate the predicted output by the model, including an almost equal number of credible and non-credible web pages. In total, 600 evaluations by 69 participants were collected for new web pages in addition to 30 web pages from the C3 dataset. This type of human validation of the output of the model for new web pages has not been done before for machine learning-based credibility classification systems, and is the first step toward investigating generalizability of the proposed approach.

6.3 Future Work

According to the P-I theory [28], the appearance and design of the web pages are among the dominant factors that people notice while evaluating the credibility of a web page. Although users aesthetics preferences are very subjective, identifying and extracting useful features that can automatically quantify usability, functionality, appearance quality, and design layout of the web page could be an exciting research line to follow.

Another possibility for future work is to find suitable ways to transfer credibility information or model predictions to users effectively and transparently. Also, since certain aspects of credibility might be more important to certain user types based on the context of the web page, building adaptively learned models that consider user preferences could be a feasible solution. In addition, an adaptive model can consider other effective factors such as user‘s knowledgeability or education level to provide more personalized credibility estimation to the user.

Due to the relatively low number of data points in the dataset, using word embeddings and state of the art deep learning models are not the ideal solution for this task. However, given enough data, using pre-trained language models and unsupervised data augmentation techniques would probably result in more accurate models.

Another limitation with the credibility assessment task is the fact that using crowdsourcing to create a high-quality dataset with a sufficiently large number of web pages is costly and time-consuming. Web pages for crowdsourcing tasks should be selected thoughtfully and cover a broad range of topics. The dataset that includes a reasonably balanced number of web pages from different credibility level is preferred but difficult to build. Therefore, using active learning or semi-supervised learning approaches could be a possible way to pursue to build more accurate and robust models with less amount of labeled data.

One important thing to consider is that content of web pages in the collected dataset, their authority, and popularity may change over time. Furthermore, sometimes the whole web page becomes inaccessible; this makes it hard to compare the performance of the developed models to the past models accurately because models have not been trained on exactly the same dataset. Moreover, perceived credibility does not necessarily match actual legitimacy of information on a website. As such, any autonomous predictive model should be revisited, redefined, and re-evaluated often to ensure that it remains accurate.

References

[1] scikit-learn: machine learning in Python scikit-learn 0.23.1 documentation. https://scikit-learn.org/stable/.

[2] Natural Language Understanding - IBM Cloud API Docs. https://www.ibm.com/cloud/watson-natural-language-understanding, May 2016. Library Catalog: cloud.ibm.com.

[3] Sonal Aggarwal and Herre Van Oostendorp. An attempt to automate the process of source evaluation. *ACEEE International Journal on Communication*, 2(2):18–20, 2011.

[4] Sonal Aggarwal, Herre Van Oostendorp, Y. Raghu Reddy, and Bipin Indurkhya. Providing web credibility assessment support. In *Proceedings of the 2014 European Conference on Cognitive Ergonomics*, page 29. ACM, 2014.

[5] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*, pages 183–194, 2008.

[6] Denise E. Agosto. Bounded rationality and satisficing in young people's Web-based decision making. *Journal of the American society for Information Science and Technology*, 53(1):16–27, 2002. Publisher: Wiley Online Library.

[7] Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. WISDOM: A Web Information Credibility Analysis Systematic. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 1–4, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[8] Alia Amin, Junte Zhang, Henriette Cramer, Lynda Hardman, and Vanessa Evers. The effects of source credibility ratings in a cultural heritage information aggregator. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 35–42. ACM, 2009.

[9] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction from the web. In *Ijcai*, volume 7, pages 2670–2676, 2007.

[10] Elisabeth Beaunoyer, Marianne Arsenault, Anna M. Lomanowska, and Matthieu J. Guitton. Understanding online health information: Evaluation, tools, and strategies. *Patient education and counseling*, 100(2):183–189, 2017. Publisher: Elsevier.

[11] Pablo Briol and Richard E. Petty. Persuasion: Insights from the self-validation hypothesis. *Advances in experimental social psychology*, 41:69–118, 2009. Publisher: Elsevier.

[12] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[13] James Caverlee and Ling Liu. Countering web spam with credibility-based link analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*, pages 157–166. ACM, 2007.

[14] John S. Caylor. Methodologies for Determining Reading Requirements of Military Occupational Specialties. 1973. Publisher: ERIC.

[15] Jeanne Chall and Edgar Dale. A formula for predicting readability. *Educational Research Bulletin*, 27:11–20, 1948.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. arXiv: 1106.1813.

[17] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[18] Thomas Chesney and Daniel KS Su. The impact of anonymity on weblog credibility. *International journal of human-computer studies*, 68(10):710–718, 2010. Publisher: Elsevier.

[19] Hal Daumé III. A course in machine learning. *Publisher, ciml. info*, 5:69, 2012.

[20] Rob Ennals, Dan Byler, John Mark Agosta, and Barbara Rosario. What is disputed on the web? In *Proceedings of the 4th workshop on Information credibility*, pages 67–74. ACM, 2010.

[21] Gnes Erkan and Dragomir R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

[22] Diego Esteves, Aniketh Janardhan Reddy, Piyush Chawla, and Jens Lehmann. Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web. *arXiv preprint arXiv:1809.00494*, 2018.

[23] Paul R. Fitzsimmons, B. D. Michael, Joane L. Hulley, and G. Orville Scott. A readability assessment of online Parkinson's disease information. *The journal of the Royal College of Physicians of Edinburgh*, 40(4):292–296, 2010.

[24] Andrew J. Flanagin and Miriam J. Metzger. *Digital media and youth: Unparalleled opportunity and unprecedented responsibility.* MacArthur Foundation Digital Media and Learning Initiative, 2008.

[25] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948. Publisher: American Psychological Association.

[26] B. J. Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, and Preeti Swani. What makes web sites credible? A report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 61–68, 2001.

[27] Brian J. Fogg. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):32, 2002.

[28] Brian J. Fogg. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*, pages 722–723, 2003.

[29] Brian J. Fogg, Cathy Soohoo, David R. Danielson, Leslie Marable, Julianne Stanford, and Ellen R. Tauber. How do users evaluate the credibility of Web sites?: a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15. ACM, 2003.

[30] Brian J. Fogg and Hsiang Tseng. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 80–87, 1999.

[31] David A Freedman. *Statistical models: theory and practice.* cambridge university press, 2009.

[32] Kris S. Freeman and Jan H. Spyridakis. Effect of contact information on the credibility of online health information. *IEEE Transactions on Professional Communication*, 52(2):152–166, 2009. Publisher: IEEE.

[33] Jerome H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. Publisher: JSTOR.

[34] Michael Gamon, Sumit Basu, Dmitriy Belenko, Danyel Fisher, Matthew Hurst, and Arnd Christian Knig. BLEWS: Using blogs to provide context for news articles. In *ICWSM*, pages 60–67, 2008.

[35] Kim Giffin. The contribution of studies of source credibility to a theory of interpersonal trust in the communication process. *Psychological bulletin*, 68(2):104, 1967. Publisher: American Psychological Association.

[36] Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. Credibility in Information Retrieval. *Foundations and Trends in Information Retrieval*, 9(5):355–475, December 2015.

[37] Katherine Del Giudice. Crowdsourcing credibility: The impact of audience feedback on Web page credibility. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–9, 2010. Publisher: Wiley Online Library.

[38] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion.* Addison-Wesley, Reading, Massachusetts, 1994.

[39] Perbinder Grewal and Swethan Alagaratnam. The quality and readability of colorectal cancer information on the internet. *International Journal of Surgery*, 11(5):410–413, 2013. Publisher: Elsevier.

[40] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st workshop on privacy and security in online social media*, pages 2–8, 2012.

[41] Manish. Gupta, Peixiang. Zhao, and Jiawei. Han. Evaluating Event Credibility on Twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, Proceedings, pages 153–164. Society for Industrial and Applied Mathematics, April 2012.

[42] Zoltn Gyngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 576–587. VLDB Endowment, 2004.

[43] Eszter Hargittai, Lindsay Fullerton, Ericka Menchen-Trevino, and Kristin Yates Thomas. Trust online: Young adults' evaluation of web content. *International journal of communication*, 4:27, 2010.

[44] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, June 2008. ISSN: 2161-4407.

[45] Paul Hitlin. *Online Rating Systems*. Pew Internet & American Life Project, 2004.

[46] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[47] Chiao-Fang Hsu, Elham Khabiri, and James Caverlee. Ranking Comments on the Social Web. In *2009 International Conference on Computational Science and Engineering*, volume 4, pages 90–97, August 2009.

[48] Melody Y. Ivory and Marti A. Hearst. Statistical profiles of highly-rated web sites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 367–374. ACM, 2002.

[49] Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, and Soyeon Park. A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 228–235, 2006.

[50] Yonggeol Jo, Minwoo Kim, and Kyungsik Han. How Do Humans Assess the Credibility on Web Blogs: Qualifying and Verifying Human Factors with Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, Glasgow, Scotland Uk, May 2019. Association for Computing Machinery.

[51] Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. Understanding and predicting Web content credibility using the Content Credibility Corpus. *Information Processing & Management*, 53(5):1043–1061, 2017.

[52] Jim Kapoun. Teaching undergrads WEB evaluation: A guide for library instruction. *C&Rl News*, 59(7):522–523, 1998.

[53] J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

[54] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999. Publisher: ACM New York, NY, USA.

[55] Donald Knuth. *The TEXbook*. Addison-Wesley, Reading, Massachusetts, 1986.

[56] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, May 2018. Google-Books-ID: nE1aDwAAQBAJ.

[57] Leslie Lamport. *LATEX — A Document Preparation System*. Addison-Wesley, Reading, Massachusetts, second edition, 1994.

[58] Leah S Larkey. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 90–95, 1998.

[59] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004. Publisher: SAGE Publications Sage UK: London, England.

[60] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[61] Alexandra List, Patricia A. Alexander, and Lori A. Stephens. Trust But Verify: Examining the Association Between Students' Sourcing Behaviors and Ratings of Text Trustworthiness. *Discourse Processes*, 54(2):83–104, February 2017.

[62] Teun Lucassen, Rienco Muilwijk, Matthijs L. Noordzij, and Jan Maarten Schraagen. Topic familiarity and information skills in online credibility evaluation. *Journal of the American Society for Information Science and Technology*, 64(2):254–264, 2013. Publisher: Wiley Online Library.

[63] G. Harry Mc Laughlin. SMOG Grading-a New Readability Formula. *Journal of Reading*, 12(8):639–646, 1969.

[64] D. Harrison McKnight and Charles J. Kacmar. Factors and effects of information credibility. In *Proceedings of the ninth international conference on Electronic commerce*, pages 423–432. ACM, 2007.

[65] Marc Meola. Chucking the checklist: A contextual approach to teaching undergraduates Web-site evaluation. *portal: Libraries and the Academy*, 4(3):331–344, 2004.

[66] Miriam Metzger. Understanding how Internet users make sense of credibility: A review of the state of our knowledge and recommendations for theory, policy, and practice. 2005. Publisher: Citeseer.

[67] Miriam J. Metzger. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American society for information science and technology*, 58(13):2078–2091, 2007. Publisher: Wiley Online Library.

[68] Miriam J. Metzger, Andrew J. Flanagin, Keren Eyal, Daisy R. Lemus, and R. McCann. Bringing the concept of credibility into the 21st century: integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication yearbook*, 27:293–335, 2003. Publisher: Lawrence Erlbaum Associates Mahwah, NJ.

[69] Miriam J. Metzger, Andrew J. Flanagin, and Lara Zwarun. College student Web use, perceptions of information credibility, and verification behavior. *Computers & Education*, 41(3):271–290, 2003. Publisher: Elsevier.

[70] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 441–450, 2012.

[71] Satoshi Nakamura, Shinji Konishi, Adam Jatowt, Hiroaki Ohshima, Hiroyuki Kondo, Taro Tezuka, Satoshi Oyama, and Katsumi Tanaka. Trustworthiness analysis of web search results. In *International Conference on Theory and Practice of Digital Libraries*, pages 38–49. Springer, 2007.

[72] Makoto Nakatani, Adam Jatowt, Hiroaki Ohshima, and Katsumi Tanaka. Quality evaluation of search results by typicality and speciality of terms extracted from

wikipedia. In *International Conference on Database Systems for Advanced Applications*, pages 570–584. Springer, 2009.

[73] Radoslaw Nielek, Aleksander Wawer, Michal Jankowski-Lorek, and Adam Wierzbicki. Temporal, cultural and thematic aspects of web credibility. In *International Conference on Social Informatics*, pages 419–428. Springer, 2013.

[74] David L. Olson and Dursun Delen. *Advanced data mining techniques*. Springer Science & Business Media, 2008.

[75] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. Web credibility: Features exploration and credibility prediction. In *European conference on information retrieval*, pages 557–568. Springer, 2013.

[76] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[77] Thanasis G. Papaioannou, Jean-Eudes Ranvier, Alexandra Olteanu, and Karl Aberer. A decentralized recommender system for effective web credibility assessment. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 704–713. ACM, 2012.

[78] Soo Young Rieh. Judgment of information quality and cognitive authority in the Web. *Journal of the American society for information science and technology*, 53(2):145–161, 2002.

[79] Soo Young Rieh and David R. Danielson. Credibility: A multidisciplinary framework. *Annual review of information science and technology*, 41(1):307–364, 2007.

[80] Victoria L. Rubin and Elizabeth D. Liddy. Assessing Credibility of Weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 187–190, 2006.

[81] Laura Sbaffi and Jennifer Rowley. Trust and credibility in web-based health information: a review and agenda for future research. *Journal of medical Internet research*, 19(6):e218, 2017. Publisher: JMIR Publications Inc., Toronto, Canada.

[82] Julia Schwarz and Meredith Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1245–1254. ACM, 2011.

[83] SEOQuake. A Powerful SEO Toolbox for your Browser SEOquake. https://www.seoquake.com/index.html.

[84] Asad Ali Shah, Sri Devi Ravana, Suraya Hamid, and Maizatul Akmar Ismail. Web credibility assessment: affecting factors and assessment techniques. *Information research*, 20(1):20–1, 2015.

[85] Ben Shneiderman. Designing trust into online experiences. *Communications of the ACM*, 43(12):57–59, 2000.

[86] Elizabeth Sillence, Pam Briggs, Lesley Fishwick, and Peter Harris. Trust and mistrust of online health sites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 663–670, 2004.

[87] Parikshit Sondhi, VG Vinod Vydiswaran, and ChengXiang Zhai. Reliability prediction of webpages in the medical domain. In *European conference on information retrieval*, pages 219–231. Springer, 2012.

[88] Julianne Stanford, Ellen R. Tauber, B. J. Fogg, and Leslie Marable. *Experts vs. online consumers: A comparative credibility study of health and finance Web sites.* Consumer Web Watch, 2002.

[89] Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. *The general inquirer: A computer approach to content analysis.* The general inquirer: A computer approach to content analysis. M.I.T. Press, Oxford, England, 1966.

[90] S. Shyam Sundar. *The MAIN model: A heuristic approach to understanding technology effects on credibility.* MacArthur Foundation Digital Media and Learning Initiative, 2008.

[91] Katsumi Tanaka, Hiroaki Ohshima, Adam Jatowt, Satoshi Nakamura, Yusuke Yamamoto, Kazutoshi Sumiya, Ryong Lee, Daisuke Kitayama, Takayuki Yumoto, and Yukiko Kawai. Evaluating credibility of web information. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, pages 1–10, 2010.

[92] Zakary L. Tormala and Richard E. Petty. Source credibility and attitude certainty: A metacognitive analysis of resistance to persuasion. *Journal of Consumer Psychology*, 14(4):427–442, 2004. Publisher: Wiley Online Library.

[93] Shawn Tseng and B. J. Fogg. Credibility and computing technology. *Communications of the ACM*, 42(5):39–44, 1999. Publisher: ACM New York, NY, USA.

[94] Vernon Turner, John F. Gantz, David Reinsel, and Stephen Minton. The digital universe of opportunities: Rich data and the increasing value of the internet of things. *IDC Analyze the Future*, 16, 2014.

[95] Lih-Wern Wang, Michael J. Miller, Michael R. Schmitt, and Frances K. Wen. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Research in Social and Administrative Pharmacy*, 9(5):503–516, 2013. Publisher: Elsevier.

[96] Aleksander Wawer, Radoslaw Nielek, and Adam Wierzbicki. Predicting webpage credibility using linguistic features. In *Proceedings of the 23rd international conference on world wide web*, pages 1135–1140. ACM, 2014.

[97] Wouter Weerkamp and Maarten De Rijke. Credibility improves topical blog post retrieval. In *Proceedings of ACL-08: HLT*, pages 923–931, 2008.

[98] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining*, pages 132–141. ACM, 2009.

[99] Minji Wu and Amlie Marian. A framework for corroborating answers from multiple web sources. *Information Systems*, 36(2):431–449, April 2011.

[100] Yusuke Yamamoto. Supporting credibility judgment in web search by Yusuke Yamamoto, with Martin Vesely as coordinator. *ACM SIGWEB Newsletter*, (Spring):1–11, 2017. Publisher: ACM New York, NY, USA.

[101] Yusuke Yamamoto and Satoshi Shimada. Can Disputed Topic Suggestion Enhance User Consideration of Information Credibility in Web Search? In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, pages 169–177, 2016.

[102] Yusuke Yamamoto and Katsumi Tanaka. Enhancing credibility judgment of web search results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1235–1244. ACM, 2011.

[103] Nazpar Yazdanfar and Alex Thomo. LINK RECOMMENDER: Collaborative-Filtering for Recommending URLs to Twitter Users. *Procedia Computer Science*, 19:412–419, January 2013.