IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 14, 2021, accepted September 2, 2021, date of publication September 9, 2021,
date of current version September 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3111527

Explainable AI for Multimodal Credibility Analysis: Case Study of Online Beauty Health (Mis)-Information

VIDISHA WAGLE[1], KULVEEN KAUR[1], POOJA KAMAT[1], SHRUTI PATIL[2], AND KETAN KOTECHA[2]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India
[2]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India

Corresponding authors: Shruti Patil (shruti.patil@sitpune.edu.in) and Ketan Kotecha (director@sitpune.edu.in)

This work was supported by the Research Support Fund (RSF) of Symbiosis International (Deemed University), Pune, India.

ABSTRACT ''One person's data or experience is another person's information'' this has become the golden rule of the 21st century which has resulted in a massive reservoir of data and immense amounts of information generation. However, there is no control over the source of this information, accessibility of this information, or the quality of it, which has given rise to the presence of ''misinformation.'' The research community has reacted by proposing frameworks and difficulties, which are helpful for (different subtasks of) recognizing misinformation. Most of these frameworks, however, fail to consider all the aspects that can contribute to making information ''credible''. Furthermore, a valid explanation for each considered feature's contribution to the model's decision stands missing in most work. With this in mind, the authors have attempted to produce a system that yields highly accurate decisions, thus effectively separating credible health blogs from their non-credible counterparts while providing valid user-friendly explanations. The study proposes an Explainable AI-assisted Multimodal Credibility Assessment System that examines the credibility of the platform where the blog is hosted, the credibility of the author of the blog and the credibility of the images that contribute to the blog. This novel framework contributes to the existing body of knowledge by assessing the credibility of misleading beauty blogs using multiple crucial modalities which would lead to an insightful information consumption by the users. The proposed pipeline was successfully implemented on multiple carefully curated datasets and correctly identified 274 non credible blogs out of 321 blogs with an accuracy of 97.5%, Precision of 0.973 & F1score of 0.986. Further, the Explainable AI model, with the help of several visualizations displayed the feature contributions for each blog & it's impact and magnitude in a concise comprehensible format. The framework can be further customized and applied to various domains where presence of misinformation is of high concern such as pharmaceutical drug information, pandemic management, financial advisories, online healthcare services and cyber frauds.

INDEX TERMS Credibility analysis, deep learning, misinformation, natural language processing, multi-modal analysis, transfer learning, explainable AI.

I. INTRODUCTION

The creation of the Internet and the advancement of the computerized age changed the correspondence scene, producing extraordinary freedoms to rapidly and effectively look for and share data, including that which identified with health. With the fast improvement of information technology and the ''big data'' approach, the receptiveness and intelligence of the recommendation system make the false score-information more conceivable to be infused. Access to information and the abundance in it has resulted in a convenience addiction, which, despite its risk quotient, is a preferred learning method. Let us consider the entire human population set. Teenagers belong to that subset, a smart generation that recognizes misinformation on the Internet and uses certain measures before consuming it. In contrast, they tend to fall prey often towards trying physical appearance-enhancing content on themselves without considering the desired outcomes. Due to the importance of ''Physical Looks'' in the internet world and the emotions attached to it, sudden changes in

The associate editor coordinating the review of this manuscript and approving it for publication was Weipeng Jing.

response protocols can make implementing any antidotes extremely difficult [1].

Unfortunately, this information is not inspected by its source, accessibility, or quality, which has caused misinformation [2]. The results of all these are concerned assaults on the recommendation system, bringing about a decline in its credibility and influencing the conclusion of the readers, making them believe in that misinformation [3]. The advanced world is loaded with such misinformation, which is creating negative and hurtful results. Fig (1) is an example of one such manipulation. The finding of such misinformation is amazingly troublesome and is regularly subject to people's abilities to get it. In this proposed work, the authors have developed an explainable AI-supported architecture that allows people to practically understand the credibility of information and helps them decide before experimenting on their bodies.

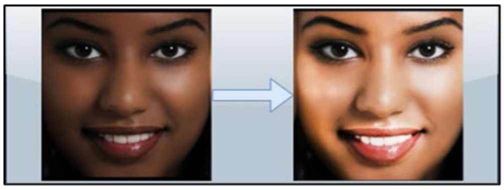


FIGURE 1. **Shows the difference between an authentic image and its doctored counterpart.**

A. IMPORTANCE OF CREDIBILITY ANALYSIS

Fake news has come to mean various things to various individuals. In this context, the term ''fake news'' is misinformation that describes news articles fabricated without reliable sources, facts, or statements. Some of these stories may play out as publicity designed to deceive the pursuer. They may be intended as misleading material for financial motivations (the author is paid every time an individual clicks the story). Recently, misinformation has increased by using web-based media since they are so readily and instantly shared. In reality, misinformed stories are only one component of the broader universe of ''fake news.''. A few stories may have a piece of truth yet come up short on any contextualizing subtleties. They may exclude any obvious facts or sources [4]. A few stories incorporate fundamental unquestionable facts yet are composed utilizing intentionally provocative language that leaves out relevant subtleties, or presents one perspective. ''Fake news'' exists inside a bigger biological system of misinformation. It is the information constructed in a structure like mainstream news to propagate made-up/misleading facts. Misinformation is false or inaccurate information that is erroneously or coincidentally made or spread; the purpose is not to misdirect [1], [5]. Disinformation is false information purposely made and spread ''to impact general assessment or dark reality'' [6]. Over the past few years, many studies aim to detect misinformation and examine its impact [7]. It is the need of the hour to educate the average reader to make them more aware while reading and responsible while sharing [8], [9].

Credibility Analysis based studies have been gaining momentum and are opening new avenues of tackling misinformation [10] [11]. ''Credibility'' is the perception of how credible an individual is based on his or her communication style, says psychologist Dan O'Keefe. In simplified terms, it assesses the degree of reliability. Thus ''Credibility Analysis'' would refer to examining this credibility quotient of a subject (blog or article or any source of information) and measuring it against a pre-defined threshold to classify it as credible (believable) or non-credible (untrustworthy). Everything from the text of the article to the pictures and even the blogger's profile can help determine its credibility. For a blog, the way the information has been portrayed, arranged & presented to the reader adds or limits its believability [12]. Thus, to detect fake news or misinformation, there can be a thorough study of the blog in its entirety. This helps provide a much broader scope to examine the degree to which a piece of information has been tampered with and detect its form (Images, Text, Platform, and so on). Using this credibility of web blogs, the detection of misinformation can be done in the early dissemination stages.

The perception of credibility defers from person to person. Fig (2) [13] represents the study of credibility types in detail. The outermost level of credibility, which relies on the most ''obvious'' cues to identify a blog as credible or non-credible, can be called construct credibility. This form utilizes the basic firsthand impressions that a blog leaves on its users. These factors are highly subjective and tend to vary very heavily from one recipient of information to another. Questions like, ''Does this blog look trustworthy?'', ''Is the information convincing?'', ''Can I believe this blog?'' often fall under this level of credibility. The most effect that it can have on a user is that it helps him/her form a bias. Examining the host platform and its features, the advertisements, and the general aesthetics of the blog (in technical detail) forms the second level of credibility. This level can be called Heuristics based Credibility. Here the user can make more informed decisions, considering that the technical details are laid out by a thorough examination of the blog & its host. Fake news, which is often masqueraded as genuine news, is more likely to be identified with the help of this level of credibility since the semantic & structural differences are highlighted more. Finally, the most complex level of credibility can be called Interaction-based Credibility, where analysis of the cues from the content, source, and images can aid in determining the authenticity of a blog. Identifying misinformation and disinformation can be attempted through this level. It focuses on a more in-depth examination of the features of a blog's credibility, providing more criteria to label a blog as credible.

B. FAKE NEWS, (MIS)INFORMATION AND (DIS)INFORMATION ON ONLINE BEAUTY HEALTH BLOGS [14]

i) A typical example of fake news could be a headline that says, ''Latest study by leading scientists reveals that the consumption of Instant Noodles can cause cancer.'' The facts

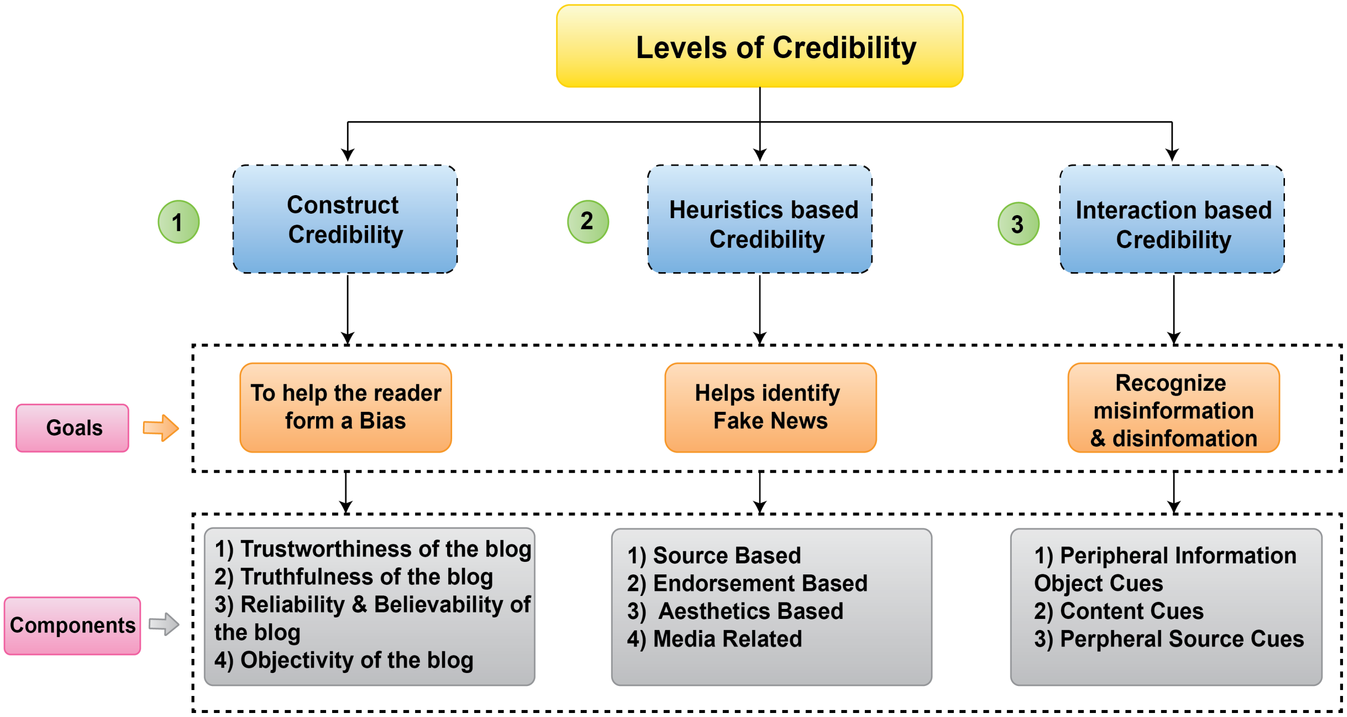

Levels of Credibility
1
Construct Credibility
2
Heuristics based Credibility
3
Interaction based Credibility
Goals
To help the reader form a Bias
Helps identify Fake News
Recognize misinformation & disinfomation
Components
1) Trustworthiness of the blog
2) Truthfulness of the blog
3) Reliability & Believability of the blog
4) Objectivity of the blog
1) Source Based
2) Endorsement Based
3) Aesthetics Based
4) Media Related
1) Peripheral Information Object Cues
2) Content Cues
3) Perpheral Source Cues

FIGURE 2. Levels of credibility.

stated are completely bogus and lack any proof, at the same time, it is perfectly dressed up as a legitimate headline. The aim in a case like this is to draw the reader's attention to terms like Scientist & Study that may induce believability.

ii) On the other hand, a teenage blogger writing something like, "Applying Baking Soda on the face can help lighten your skin tone," becomes an example of misinformation. Here, the blogger aims not to mislead or deliberately cause harm, but due to lack of expertise & knowledge, they believe in such information to be true and pass it on to their readers.

iii) A blog that is probably sponsored by a skin whitening beauty product, suggesting that "Applying a particular cream for seven days produces visible results", deliberately misleads the readers by using fake studies or doctored images to increase their sales. This becomes an example of disinformation.

To a naive reader, the difference between these three may not be evident; however, their effects vary significantly in intensity. Fig (3) shows a taxonomy of fake news.

C. MOTIVATION

The web has become a well-known asset for health sources that provide remedies & suggestions which are experimented with, especially among teenagers [15], [16]. Nonetheless, individuals can undoubtedly get misinformed given the huge measure of inaccurate data on the web. For instance, the idea that eating apricot seeds fixes cancer is a misguided judgment that can be discovered on the web. There is no logical proof to help the case; indeed, it is grounded that eating apricot seeds may cause cyanide poisoning. People have consistently gotten data from outside the conventional medical services framework, and wellbeing falsehood and disinformation are not new. Seeing how the web has changed our commitment to wellbeing, (mis)information and whether people can effectively assess integrity is a significant assignment. This is because falsehood concerning health has especially severe results in individuals' satisfaction and, surprisingly, their danger of mortality. As a result, controlling the spread of misinformation while fostering trust in the entire news ecosystem has become vital.

D. CHALLENGES

However, detecting fake news/ misinformation presents unique challenges.

1. Even though deep learning-based misinformation detection methods have been successful, most of these methods mainly focus on detecting a particular form of credibility. However, to measure the reliability of an article most accurately, different types of credibility assessment techniques are needed too.
2. The second issue is that when a user discovers that a website is fraudulent and lacks trust, he has no idea what makes it fake. Explaining why a website was discovered to be fraudulent is desirable because it can provide fresh experiences and knowledge that was previously unknown, even to experts.

E. PROPOSED SOLUTION

The authors propose to isolate the validity into three sections (1) Web (2) Author and (3) Image. An ingenuine website

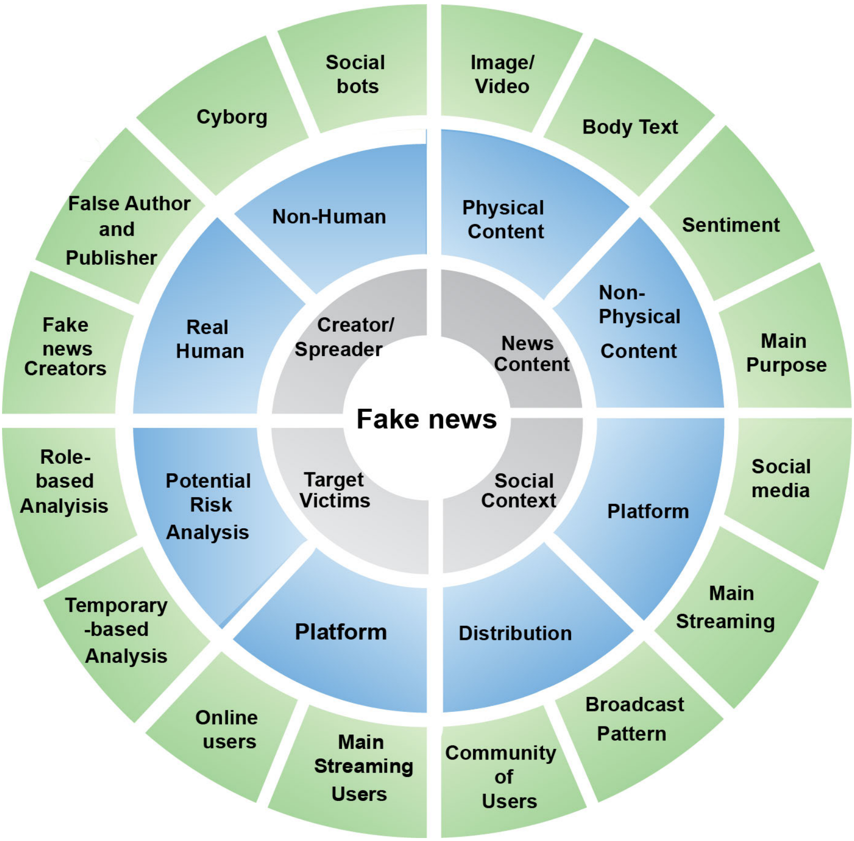


FIGURE 3. Analyzing fake news.

is more likely to host articles that could be invalid or not pertinent to the content promised. On the other hand, regardless of whether the site is credible, thinking about the author's data (domain mastery, article composition style) is vital as even on trustworthy sites, individuals with less knowledge can spread bogus information. Lastly, a lot of the articles present pictures to attract clients. But many of these pictures are photoshopped utilizing numerous strategies. This study presents credibility scores for each part through a Multimodal Credibility Assessment System [17]–[20] and then finally an X-AI System [21], which explains the factors contributing to the model's decisions.

F. PAPER ORGANIZATION

This paper has nine sections to follow. Related work section (Section II), showcases a thorough literature review of existing credibility assessment techniques & explainable AI techniques for fake news/misinformation detection. Following this, the dataset organization section (Section III) describes the methods used for the preparation of each dataset. The next two sections namely, Multimodal Analysis (Section IV) and Explainable AI (Section V) represents a total of four major implementation pipelines with a detailed explanation of the algorithms and techniques used in each. The next section i.e. Results (Section VI) is divided into two modules 1) Sub-Modules results 2) System Results. Finally the paper includes a discussion section (Section VII) that elaborates the challenges and limitations of this project. A brief note on the future scope of the project is discussed in the Future Work section (Section VIII). This is followed by the Conclusion (Section IX) and References (Section X) sections.

II. RELATED WORK

Internet is a huge wellspring of data covering different scope of points. From recent developments to education to healthcare, each sort of detail is composed, perused, or shared by a great many clients all over the globe. Web Health web journals or Wellness websites contribute significantly to online health information, generally by giving tips and solutions for afflictions. Few sites additionally address significant medical problems. It gets imperative in such cases to forestall any spread of deception as it might prompt unfortunate and serious outcomes.

A. CREDIBILITY ASSESSMENT

Credibility as a quality is formed by considering various estimations. It does not exclusively rely upon the source or the content and is achieved from a variety of measurements. Credibility is also synonymous with believability [22], and similarly, it can be associated with various other measures like

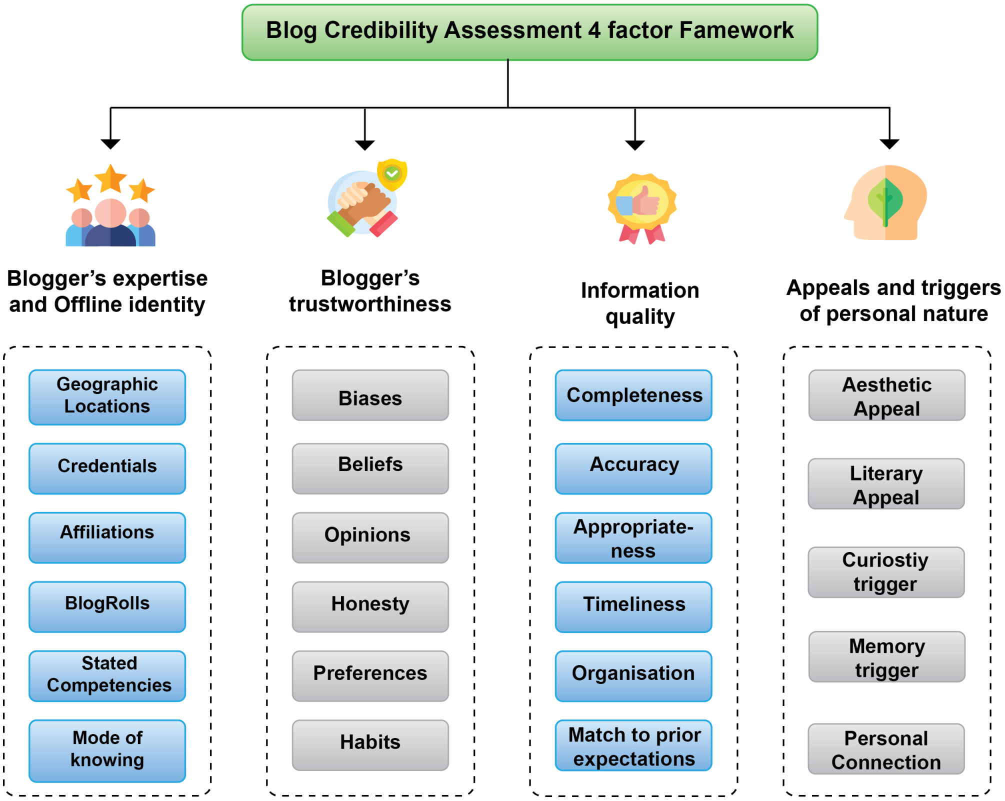


FIGURE 4. Blog credibility assessment 4 factor framework.

quality, authority, popularity, and persuasion. Credibility can be classified into various categories depending on the subject under assessment [23]. Source, media, and content credibility are the most common assessment techniques which examine the actual data/information.

One such technique is web credibility, which relates to the credibility of the various blogs and articles available in massive quantities without filters across the world wide web [24]. Burbles [25] stressed the importance of a separate accessibility framework for the web due to its complex link structure, speed, and abundant features. The web credibility project by Stanford [23] expresses credibility evaluation as a 5-point multidisciplinary framework that identifies the contributions of various features towards a site's credibility. Web blogs, in particular, serve to be a perfect venue for credibility assessment owing to their wide availability and public nature. Rubin and Liddy [26] proposed an analytical 4 factor framework to automate credibility assessment of the web blogs. Fig (4) depicts this framework.

Web Blogs' credibility can be assessed using famous traditional algorithms like PageRank [27] and HITS [28]. A web blog could fall prey to content spam - spamming of title, body, meta tags which can be checked by examining TF-IDF scores [29], link spam - to increase web authority scores can be detected by algorithms like graph regularization [30] and link pruning [31] and other spams like cloaking, redirection which can be detected by tracking user behaviors [32].

Extensive research has been done on the content of web blogs to determine credibility based on different factors. Credibility signals can be identified in the author's sentiments, the expertise reflected in the content, the readability, grammar, and vocabulary used. It was observed that more words, sentences, numbers reflected more in accurate information blogs than in comparison to misinformation blogs [33]. Linear Regression & Neural Networks were the two approaches suggested for web page credibility by Jaworski in 2014 [34]. In 2009, work towards credibility assessment of Arabic blogs [35] categorized features as blog level (presence of Author name, number of comments, etc.) and post level (spelling, emoticons, punctuations, etc.) and labeled each blog into three categories: *Credible, Not Credible, Questionable*. As an extension to this approach, to solve the scarcity issues of Arabic web-blogs, a deep co-learning approach was proposed [36]. It was observed that the SVM model had an F1-score of 0.57, whereas this model increased the score to 0.63. R. Manjula and M. S. Vijaya suggested a predictive model based on deep neural networks using an elaborate dataset of health pages. This model achieved great results [37]. Furthermore, K. Popat, in 2019, suggested analysis of the credibility by making use of textual claims. For each claim in the blog, it was cross checked with multiple relevant web sources, and then analyzed independently to estimate an opinion for each [38]. Another study made use of cosine similarity score to classify blogs into various

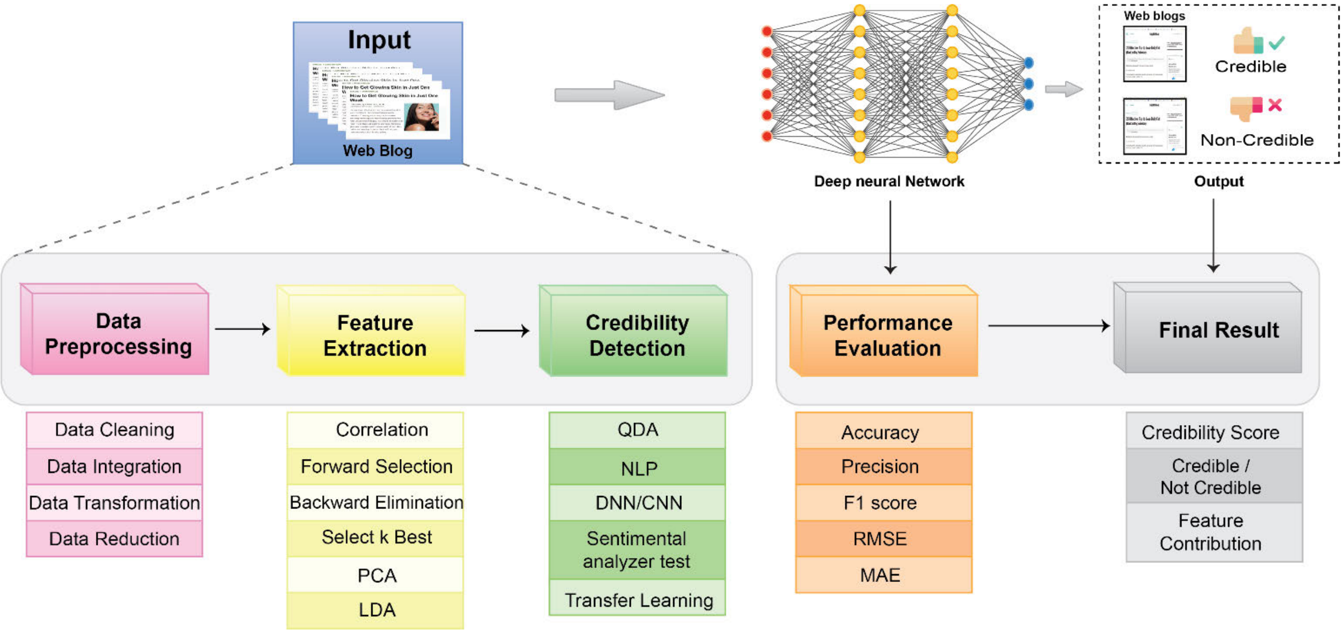


FIGURE 5. Typical credibility assessment architecture.

categories, the system used web crawling to extract features like title, date, author, content, language, tags, and permanent link to determine credibility [39].

Several blogs have multiple images along with the content. These images are often advertisements/spam that does not contribute to the quality of the blog [40]. Jelena Kocic and Branko Livada in 2016 gave 6 parameters for image quality assessment [41]. Very limited research has been conducted to evaluate the credibility and quality of these images. In 2012, a study proposed Image Texture Analysis Based Image Spam filtering [42]. Low-level image features were used for classification with classifiers like SVM, and RF was used to classify images as credible/not credible. The recall, average precision, and accuracy were observed to be 98.6%. In 2019 Andreas Rossler[1] Davide Cozzolino[2] Luisa Verdoliva[2] Christian Riess[3] Justus Thies[1] Matthias Nießner[1] generated a large-scale dataset of manipulations by classical computer graphics-based methods like Face2Face and FaceSwap as well as learning-based approaches like DeepFakes and NeuralTextures. It gave an accuracy of 70.10% [43]. Another study used an Ada-Boost-like transfer learning algorithm to classify fake images, a CNN model that exceeded performance in terms of several baselines [44].

In their study of image credibility, Yusuke Yamamoto and Katsumi Tanaka tried to establish "supportive" relationships between images and text of a blog [45]. Their work proves that blogs with more of such image-text pairs are more credible. They tend to focus on relevant pictures & text more, thus having fewer chances of diverging from the subject and disseminating misinformation. Another study proposed a framework, MediEval 2015 [46], a semi-automated approach that examined viral images and sent updates to journalists in real-time. The images were scraped off tweets on various bogus topics. This approach, too, utilized image credibility analysis along with text credibility to detect misinformation. While image credibility evaluation by a user is purely subjective, which means it relies heavily on a user's ability to trust or their loyalty to a platform or source of news, other factors may influence their decisions. A study proved that factors like a user's own knowledge of image doctoring techniques, social media presence, and general use of the internet and various other tools were more influential than any other factors [47]. Fig (5) depicts a typical credibility assessment system with different techniques. Table 1 shows the previous work done in Credibility Analysis.

After a thorough review of the findings in credibility assessments of web blogs, a common pattern is observed in their limitations. One of which was, assessing credibility in only specific modalities. Most work has been targeted towards examining credibility in a single modality, like only text or images. To ensure a complete examination of the blog, it is vital to inspect all the features that contribute to its credibility. Multimodal credibility thus provides better avenues to detect misinformation by inspecting several different types of factors that can affect the credibility of a blog. Another limitation was the lack of explanations for the outcomes. While most papers have successfully classified blogs as credible/ non-credible, the recipient of this information is often unaware of what features contributed to this result. It can be advantageous for a reader to be aware of what aspect of the blog makes it more reliable & believable and which aspect is doctored or undependable. Explainable AI can prove to be a solution here by explaining how each feature

TABLE 1. Work on credibility analysis.

Papers	Modalities			Algorithm used	Methodology Used	Application Area	Dataset Used
	Text	Image	Web				
Popat et al. (2019)[38]	✓	✗	✗	Ensemble of layers in Neural Networks (DNN, LSTM, CNN)	This approach incorporated external evidence or counter-evidence for each claim in the article.	Political News	Scopes and PolitiFact datasets. ***Accuracy: 78.93%***
Choudhary et al. (2019)[48]	✓	✗	✗	Quadratic Discriminant Analysis (QDA), Forests (ensemble of Decision Trees) AdaBoost Classifier, Multi-Layer Perceptron Neural Network (MLP-NN)	This paper proposes a novel approach called CREDO. This approach captures the various features of an article that can help determine its credibility and feeds them to a neural network that learns each of their contributions [48]	Analysis of Unstructured text articles in an open-domain setting.	Snopes Dataset, SemEval 2016 Dataset: ***Accuracy 83.3%***
Rubin et al. (2006)[26]	✓	✗	✗	NLP analysis of text	This paper proposed a 4-factor model framework to help determine the factors contributing to an article's credibility.	Web Blogs	-
Situala et al. (2019)[33]	✓	✗	✓	Co-authorship networks, Sentiment analyzer - VADAR, Shapiro-Wilk test, Mann Whitney U test, Flesch-Kincaid reading-ease test	This work focused on credibility based on content & author details. The authors proposed a concept of co-authorship networks; Various content-related credibility aspects were also explored.	Fake News	PolitiFact, BuzzFeed ***Precision and recall between 0.7-0.8 and average F1-score 0.80***
Helwe et al. (2019)[36]	✓	✗	✓	Deep Co-Learning	The semi-supervised deep co-learning method proposed by this paper outperforms traditional models like SVM TF-IDF (0.57) & Ensemble CNN (0.50)	Trending News (Arabic language)	Labeled Arabic Web blogs Dataset. ***F1 score: 0.63***
Olteanu et al. (2013)[49]	✓	✗	✓	Decision Trees, , Naive Bayes, Extremely randomized trees (ERT) for classification. SVM and ERT's variants for regression.	The paper presented a super-set of 37 features vital to determine credibility.[49]	Web Blogs	Public Blog Features Dataset by Microsoft Accuracy: 75% (classification), Regression ***Improvement of approximately 53% was recorded on RMSE and MAE.***
Juffinger et al. (2009)[39]	✓	✗	✗	NLP- Centroid Cosine Similarity	This paper introduces an additional dimension for credibility assessment - quantity structure. Each blog is analyzed and compared with a news corpus, and hence a credibility level per blog is generated	Trending News	German APA News Corpus, Mined Text from Wikipedia (German, English, French) ***Precision of 0.73***

contributes (positively or negatively) to a blog's credibility and highlighting the magnitude of its effect [50].

B. EXPLAINABLE AI FOR FAKE NEWS/MISINFORMATION ANALYSIS

There has been a surge in AI-based products and services to process large data, enable autonomy, and enhance the end-users experience in recent years. However, a lack of transparency and specific ability in advanced AI algorithms can potentially result in unfair and unsafe decision-making. Explanation of factors contributing to model decisions, can add to urban thinking by improving our capacity to assess computerized substance and arrive at supported resolutions. Answers to questions like, how difficult is the detection? Are all features needed by the model for prediction? How much does a feature contribute to making a news fake? were necessary to enhance user's understandability [51]–[53].

All these questions were answered by designing an AI-based model [54], [55] and running human-subject experiments on hypothesis testing [56]. Many techniques were used that includes RNN, GRU [55], and BI-LSTM [56] but describing how total importance should be distributed among the features was one of the successful methods used in 2009 [57]. Fan Yang, Shiva K. Pentyala, and many others designed and implemented an XAI system, XFake, based on three frameworks (i.e., MIMIC, ATTN and PERT). By understanding the outputs achieved from these frameworks, they could derive appropriate explanations for interpreting detection outcomes [54]. In 2009, the SHAP (for Shapley Additive Explanations) method was analyzed to explain a model by Julio C. S. Reis and André Correia with many others [57].

Most of the explanations were proposed to derive explanation from the perspectives of news contents and user comments [57], for which many researchers used

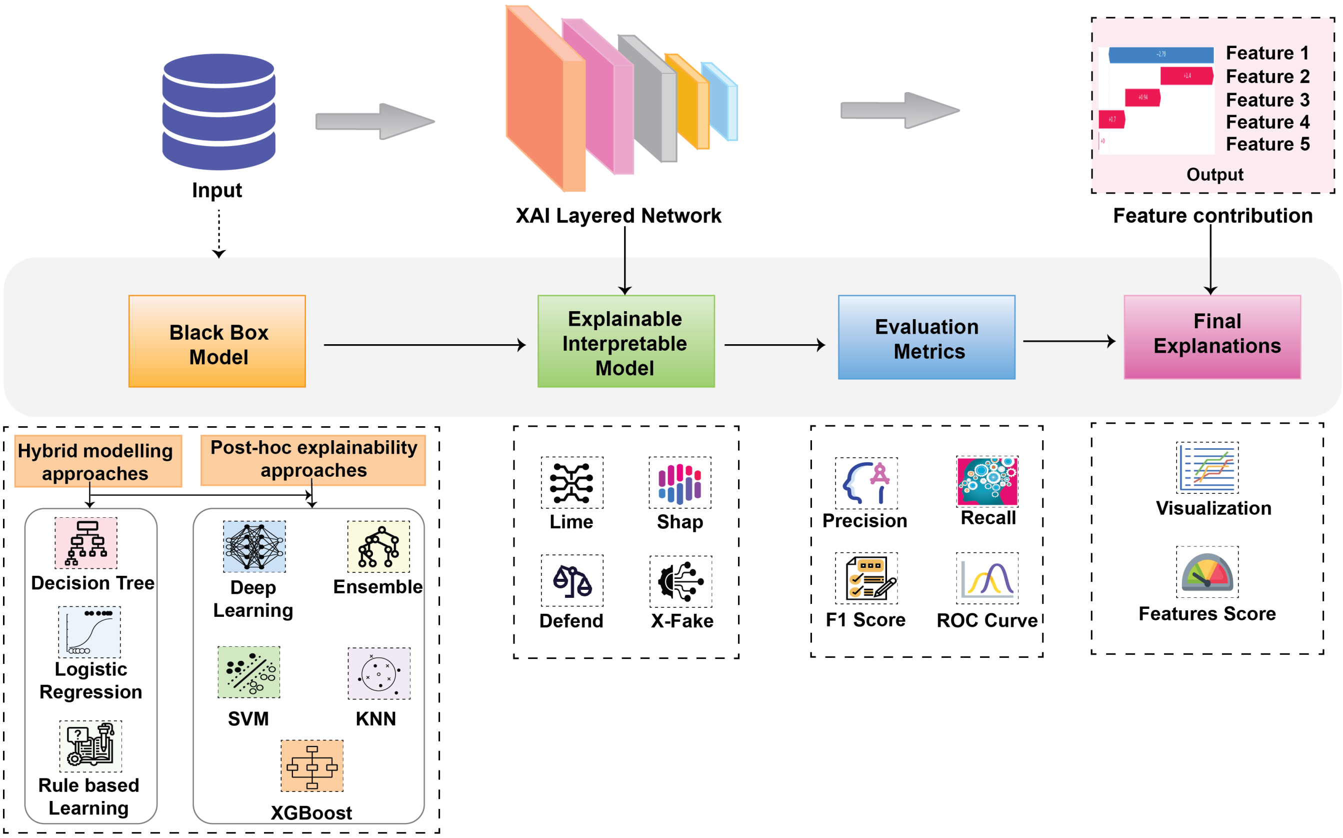


FIGURE 6. Explainable AI for credibility analysis.

Graph-aware Co-Attention Networks [60]. Kai Shu and Limeng Cui, along with others proposed a framework, named as dEFEND that consisted of four major components (1) a news content encoder component, (2) a user comment encoder component, (3) a sentence-comment co-attention component, and (4) a fake news prediction component [57]. While these researchers focused on different models for explainability, Mingxuan Chen and Ning Wang defined ''influence scores,'' which they used to measure the influence of various types of features on the final decision [61]. Fig (6) shows an XAI-assisted architecture for Fake news / Misinformation Analysis. Table 2 shows work done Explainable AI for Fake news / Misinformation Analysis.

III. DATASET PREPARATION

To start with the implementation, 300+ blogs (webpages) were handpicked, covering various health-related content available online. From beauty & lifestyle to health & skincare, the authors considered websites offering several tips & remedies to their users, making it vital to assess their credibility. Our proposed framework divides credibility into three major modules. To assess the credibility of each kind, a customized dataset was prepared, on which each method was applied, resulting in a score for each sub-module. The scores generated were entered into a master database, further subjected to the X-AI model, presenting its interpretations in an understandable user format. Each feature (label) considered for every dataset affects the overall credibility of the web-page. Table 3 provides a summary of the datasets and Table 4 explains the labels and their meanings for each dataset considered. Fig (7) explains the approach in a simplified way, explaining the contribution of each dataset & the score evaluated using it.

A. SOURCE CREDIBILITY ASSESSMENT

The host platform is essential when examining the reliability of a blog. Various features contribute to making a platform popular & more dependable for users. Our approach considers 14 of the most important features of a web blog that make it credible. These features include Internal Links, Meta-Tags Page Titles, URL-Format, Amount of Content, Popularity, Freshness, Twitter, Images, Printability, Server Behavior, Analytics, Headings & Mobile. These features were evaluated on a scale of 10 and were examined using free SEO testing tools. A non-weighted average of all these contributors provided us with an overall score of 10 for each web blog. For the purpose of our project, we made use of Nibbler, a free tool available online to help examine the performance of any website. The tool evaluates a website based on a number of parameters like Popularity, Accessibility, Security as well as Social Media integrations. This helps provide a measure of

TABLE 2. Work on explainable AI for fake news / misinformation analysis.

Research paper	Dataset used	Features used	The algorithm used & Accuracy obtained	Research findings
Chen et al. (2020) [58]	PHEME, RumorEval2019 dataset	Twitter features (Follower's count, Account age, Verified or not, etc.)	Universal sentence embedding (USE), the stochastic gradient descent + momentum (SGD + Momentum), 6-layer multi-head attention (MHA) ***F1 score = 0.453 Accuracy = 0.559***	This paper describes an architecture for rumor detection using three classes of features obtained from language embeddings.
Yang et al. (2019) [54]	PolitiFact Dataset	News Statements (text)	ATTN, MIMIC, PERT, XG Boost Accuracy ***1.MIMIC = 67.1%*** ***2.ATTN = 67.3%*** ***3.PERT = 53.2%***	Their XFake system, which assists users in identifying the legitimacy of news stories, has been presented as an explanation
Mohseni et al. (2020) [56]	news dataset of news headlines in Snopes	Textual features	LSTM network with a self-attention layer trained on news headlines, Bidirectional LSTM, Pearson Chi-square test ***Accuracy = 66.6%***	They designed a news review interface with a built-in AI assistant and then ran human-subject experiments for hypothesis testing.
Reis et al. (2019) [57]	Buzz Face Dataset	Language Structures, Lexical Features	XGBoost, Shapley, ensemble technique ***AUC values are higher than 0.85*** (0.882±0.024 of AUC).	They found the distribution of factors in terms of variability and accuracy. Using Shape, they found a negative and positive impact of the features on the output value.
Lu et al. (2020) [59]	Twitter15, Twitter16 dataset	Tweets features (textual + numeric)	GRU-based Representation, CNN based representation, Dual Co-attention Mechanism ***Accuracy = 16%***	They developed Graph-aware CoAttention Networks (GCANs) and provided reasonable explanations for the trends observed.
Hsu et al. (2020) [40]	Gossip cop, PolitiFact	News contents, user comments (Textual Features)	RNN, GRU NDCG =28.2% ***Precision = 30.7% Accuracy = 90.4%***	To figure out why a news article is fraudulent, an algorithm is being developed that uses news content and user comments.

TABLE 3. Overall dataset information.

DATASET PARAMETERS	*VALUES*
TOTAL NUMBER OF BLOGS UNDER STUDY	320
Average number of words per blog text	1033
Average images considered per blog	2
Pages considered per blog	According to the Blog Length
Language Considered	English
Average Typo Free Text %	91.66%
Average Readability (As per Flesch Scale)	45.93% : Low Readability High Difficulty
Average Grammar Proficiency (CFG Grammar)	~98%

overall performance for a web-page and highlights the areas that accelerate/decrease reliability.

Each feature chosen here was selected based on the amount of impact it leaves on a blog visitor. Factors like the popularity of a blog are determined by the number of fresh viewers that access it daily while also considering its ranking on the Google Search Index. A site with meta-tags has a better GSI ranking. Freshness measures how often the writer updates a blog. A frequently updated blog is more reliable, as it is more likely to accommodate any new findings & studies. Analytics & Server Behaviors are qualities heavily reliant on the platform; low latency & faster response time, along with customized advertisements, make the experience better. A responsive web blog (one that retains its design on mobile & desktop) & compatible printing resolutions have a higher Mobile & Printability score. The amount of content & images also contributes to credibility. A higher ratio of images to text or vice versa proves to be less effective when compared to an equalized distribution. Other features like a good URL format & a substantial amount of internal links contribute equally to the overall platform credibility score.

PageRank analyzes directed graphs and, in particular, the web link structure. It creates the impression that a user browses a site through links in random order. Rankings are derived from a link graph, which considers how each web-page is linked to the next. The extraction uses the Beautifulsoup library in Python to pull data out of the HTML and XML documents. It extracts the weblinks and their values and then stores them in the form of a matrix. Each page gets exactly 1 "vote," which is further reduced by the number of outgoing links the page has. For example, if Page A has ten outgoing links, each outgoing link counts as 0.1 vote which when passed through the page rank algorithm gives the relevant page rank score.

TABLE 4. Information per dataset.

calculated using SEO Optimizing tools like Nibbler	Mobile	Is the website optimized for viewing on a mobile/tablet?	This ensures good user experience regardless of the medium used to access the web page
	Headings	Are the headings well defined for search engines and visitors?	This is to ensure that the viewer lands on the appropriate site for the content they've requested
	Page Titles	Does the page have well defined titles?	This is necessary to ensure clear knowledge consumption for the viewers
	URL format	Are appropriate web addresses used throughout the page?	A uniform format should be ensured for all the weblinks referred on the page, this makes the webpage more trustworthy.
	Amount of Content	Does the word distribution correlate to its search engine ranking?	A webpage should not have too little/too much content to display.
	Popularity	What is the current ranking of this website? Has there been a decrease in popularity?	A consistently popular webpage is more likely to be credible (as it is used more)
	Freshness	Is the website updated frequently?	This ensures that constant work is being done on the webpage, and the content isn't outdated.
	Twitter	Is the website linked to a twitter account? How are the followers/tweets made by the account?	Seeing the performance of a webpage's social media could also contribute to its popularity.
	Printability	Are the pages optimized for printing?	In case a user needs the content of the page for quick reference, the page should be suitable for printing.
	Meta Tags	Does the page have its meta tag descriptions in place?	Meta-tags are used by search engines for indexing and ranking, a check on their presence is thus necessary.
	Server Behavior	Does the server handle missing pages? Does it follow a good encoding practice to reduce load time?	A good web development practice is to ensure the presence of error pages; thus, their presence should be checked for.
	Analytics	Does the website use any analytics software?	Webpages that utilize analytics might be more likely to incorporate more of the users' preferences/needs.
	Overall	Cumulative average of each of these scores	This helps provide a measure of the overall webpage credibility for this project.
Author	Blog Text	Text scraped off the blog	To analyze the writing style of the author it becomes important to examine the text.
	No of words	Number of words in the text	This helps check how concise/lengthy the blog content is.
	Grammar Score	Score obtained after applying grammar test	This helps provide a measure of how proficient the grammar is of the blog
	Readability Score	Score obtained after applying readability test	This helps provide a measure of how readable the text is of the blog
	Domain Score	Score obtained after applying domain expertise test	This helps provide a measure of how domain efficient the blog is
	Typo Score	Score obtained after applying typo detection test	This helps provide a measure of how free the text of the blog is of spelling errors.
	Author Score	Cumulative average of each of these scores	This helps provide a measure of the overall author credibility for this project.
Images	Credible Images	Set of images perceived as credible	In order to check for credibility of images, it was necessary to examine them.
	Non-Credible Images	Set of images perceived as not credible	In order to check for credibility of images, it was necessary to examine them.
Page Rank	Serial Number	No of website	This uniquely identified a web-page
	Page 1	Page rank score of 1st web page of each website	This score resulted after applying the algorithm on the first page of the website under consideration
	Page 2	Page rank score of 2nd web page of each website	This score resulted after applying the algorithm on the second page of the website under consideration
	Page 3	Page rank score of 3rd web page of each website	This score resulted after applying the algorithm on the third page of the website under consideration
	Page 4	Page rank score of 4th web page of each website	This score resulted after applying the algorithm on the fourth page of the website under consideration
	Page 5	Page rank score of 5th web page of each website	This score resulted after applying the algorithm on the fifth page of the website under consideration
	Desired Page Score	Page rank score of desired i.e 5th web page of each website	This score resulted after applying the algorithm on the fifth page of the website under consideration
	Scaled	Page rank score scaled to 1-10	The final page rank score is scaled for easy use

TABLE 4. ***(Continued.)*** **Information per dataset.**

Master Dataset	Desired Page Score	Page rank score of desired i.e 5th web page of each website	This score resulted after applying the algorithm on the fifth page of the website under consideration
	Scaled	Page rank score scaled to 1-10	The final page rank score is scaled for easy use in this project
	Final	Final page rank score rounded up to 2 decimal places	We finally round off the score to 2 decimal places
	Blog	Link of the blog	Each website is identified by its unique link
	Website_Score	Score obtained after website behavior analysis	This helps provide a measure of the overall webpage credibility for this project.
	Page_rank_score	Score obtained after page rank analysis	This helps provide a measure of the overall page rank score for this project.
	Author_Score	Score obtained after author credibility analysis	This helps provide a measure of the overall author credibility for this project.
	Image_Quality_score	Score obtained after image quality analysis	This helps provide a measure of the overall image quality for this project.
	Image_Score	Score obtained after image credibility analysis	This helps provide a measure of the overall image credibility for this project.
	Final_Score	Cumulative average of each of these scores	This helps provide a measure of the overall credibility for this project.
	Credible/Not	Classifying image as credible/not as per the final score obtained	This helps classify a blog as credible/not credible as per the threshold set for this project

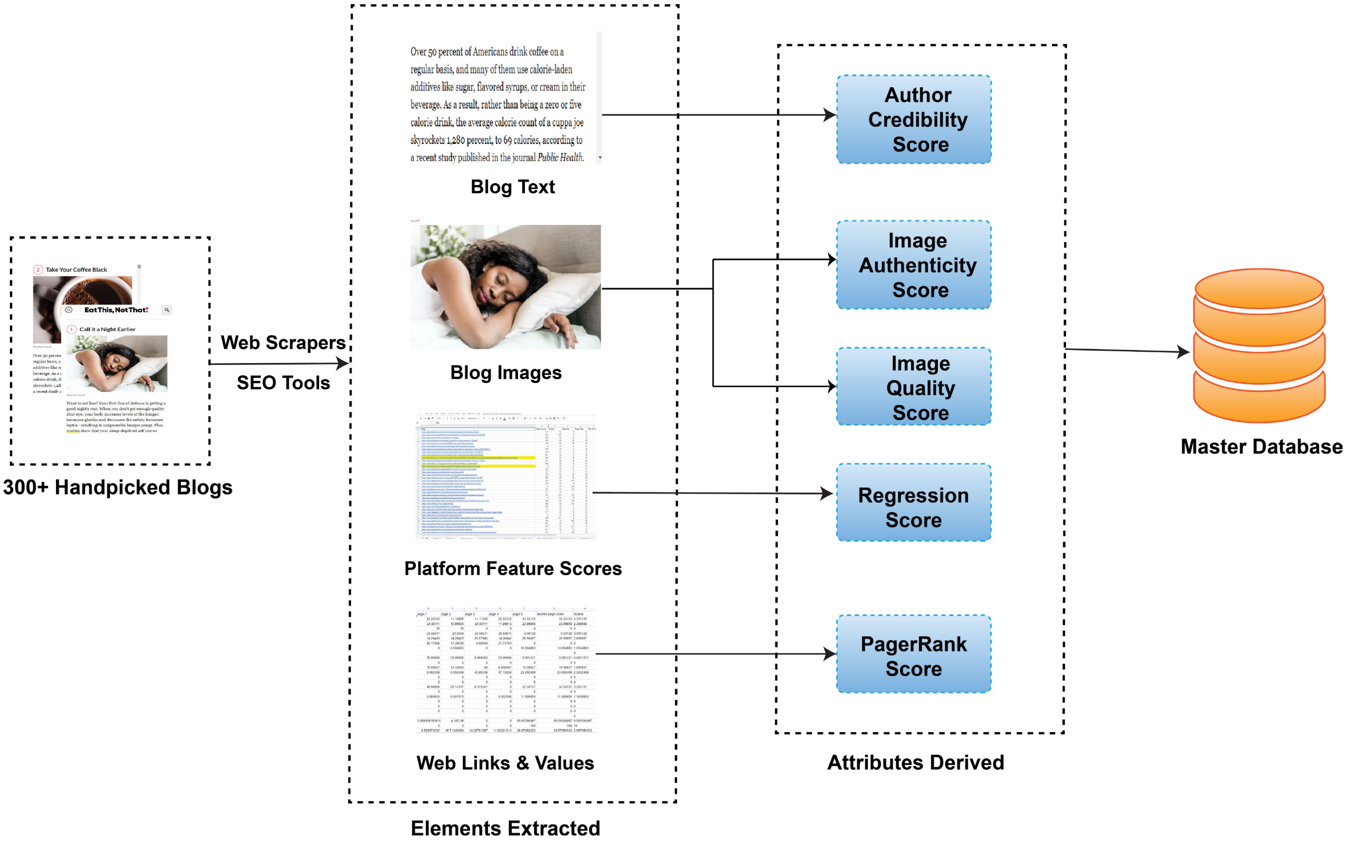

Over 50 percent of Americans drink coffee on a regular basis, and many of them use calorie-laden additives like sugar, flavored syrups, or cream in their beverage. As a result, rather than being a zero or five calorie drink, the average calorie count of a cuppa joe skyrockets 1,280 percent, to 69 calories, according to a recent study published in the journal *Public Health*.

FIGURE 7. **Dataset preparation.**

B. AUTHOR CREDIBILITY ASSESSMENT

Author authority is a compelling factor in blogging. It not only determines the amount of traffic a site attracts but also affects the impact it creates. Loyal followers of authors tend to visit a variety of platforms to be able to consume their content. But how can the credibility of one such author be determined? While liking someone's content is subjective, it becomes vital to decide based on how genuine the information is in health blogs. The writing style of an author depicted by the text of the blog can display signs of credibility. Typos, bad grammar,

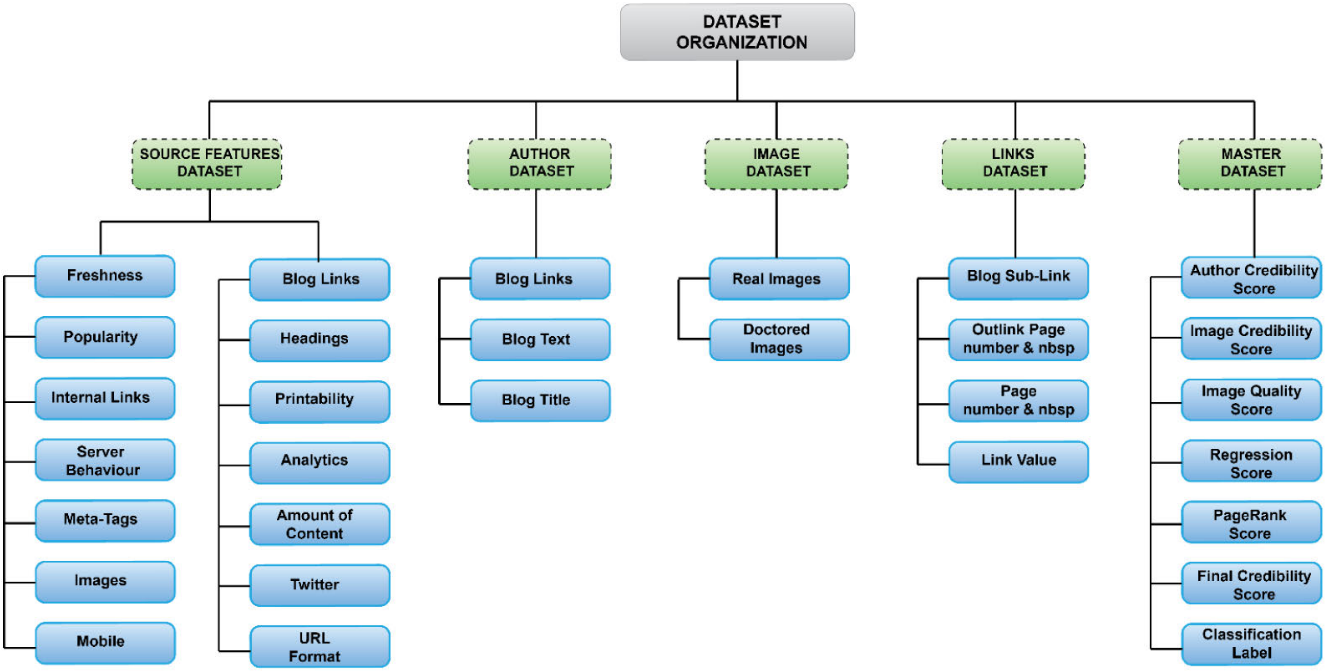


FIGURE 8. Dataset organization.

lack of technical terms, and unnecessary special characters make the content less reliable. The blog text can be analyzed to extract all these features. Beautifulsoup is used to scrape this content off the shortlisted blogs.

C. IMAGE CREDIBILITY ASSESSMENT

To compel more audiences into following their tips/remedies, blogs tend to use doctored/misleading images as a part of their platform. These images are usually photoshopped versions of their originals that show glorified versions of the results the blog claims to produce. A web scraper is built, using beautiful soup, to extract images of these websites. Each doctored image was labeled as "Fake," and the other was named "True." This labeling was done to the best of the knowledge. This dataset of 800+ images was used to get the quality & credibility score for each image.

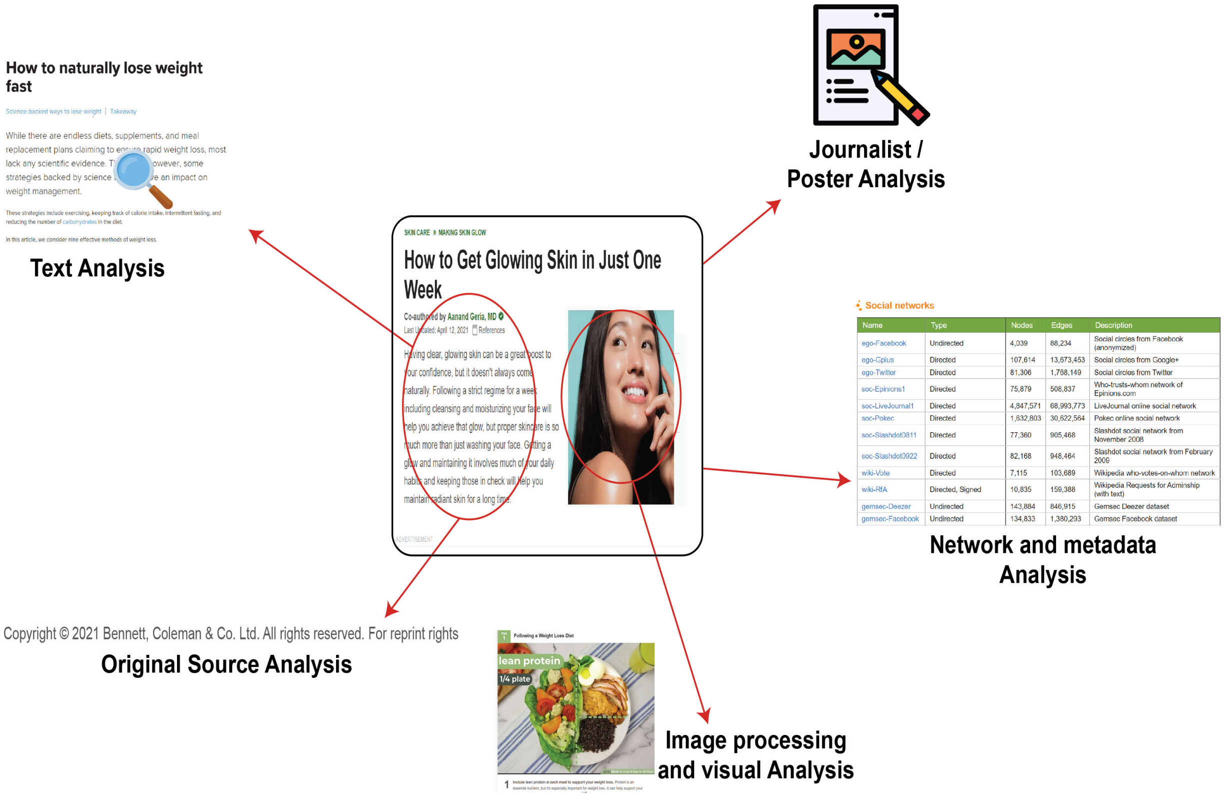
D. DEVELOPING THE MASTER DATASET

After performing Multimodal Credibility Analysis on the selected web-blogs, a score for each type of analysis is calculated. The analysis techniques and derivation of scores for each modality have been described in this paper's later sections. All scores are added to this master dataset. determine the overall credibility score of each blog, which is calculated as the non-weighted average of these individual analysis scores. After setting a threshold for credibility each blog is classified as credible or non-credible and provided with a value of 0 or 1 for the same, respectively. The XAI model uses this dataset to make suitable predictions and provide the final system results. Fig (8) depicts the dataset organization for this project.

IV. MULTIMODAL CREDIBILITY ANALYSIS

Aside from the blog's actual content, its images, design, layout, and structure all play a role in determining the first impression it leaves on its visitors. It is a blog's visual, aesthetic, or even literary appeal that can entice a reader. While the majority of these are subjective to his preferences, some of them can be measured. In this study, the authors investigate the platform, its images, and text. The platform on which the blog is hosted can have various features that help increase its relevance and credibility. Some of the factors contributing to a platform's credibility are how frequently the blog is updated and how easily it is accessible. A well-connected platform is also more likely to gain the trust and loyalty of its readers.

On the other hand, the blog's content also has a long-lasting impact on the reader, and it is through this text, the author can attempt to establish a personal connection with them. Finally, it is critical to determine whether the blog's images are correct and relevant. Visitors rarely prefer unnecessary low-quality images that promote bogus products or narrate doctored results/techniques. Therefore, multiple factors come into play while determining the blog's dependability & authenticity. Relying solely on one of these factors would only provide a skewed picture of a blog's credibility. Thus, Multimodal credibility [60], [61], in which the study examines more than one type of data to determine the blog's overall quality, is best suited to the goals of this project. Fig (9) shows the multimodal analysis of a blog.



How to naturally lose weight fast

Science-backed ways to lose weight | Takeaway

While there are endless diets, supplements, and meal replacement plans claiming to ens... rapid weight loss, most lack any scientific evidence. T... owever, some strategies backed by science ... ve an impact on weight management.

These strategies include exercising, keeping track of calorie intake, intermittent fasting, and reducing the number of carbohydrates in the diet.

In this article, we consider nine effective methods of weight loss.

SKIN CARE » MAKING SKIN GLOW

How to Get Glowing Skin in Just One Week

Co-authored by Aanand Geria, MD
Last Updated: April 12, 2021 References

Having clear, glowing skin can be a great boost to your confidence, but it doesn't always come naturally. Following a strict regime for a week including cleansing and moisturizing your face will help you achieve that glow, but proper skincare is so much more than just washing your face. Getting a glow and maintaining it involves much of your daily habits and keeping those in check will help you maintain radiant skin for a long time.

ADVERTISEMENT

Social networks

Name	Type	Nodes	Edges	Description
ego-Facebook	Undirected	4,039	88,234	Social circles from Facebook (anonymized)
ego-Gplus	Directed	107,614	13,673,453	Social circles from Google+
ego-Twitter	Directed	81,306	1,768,149	Social circles from Twitter
soc-Epinions1	Directed	75,879	508,837	Who-trusts-whom network of Epinions.com
soc-LiveJournal1	Directed	4,847,571	68,993,773	LiveJournal online social network
soc-Pokec	Directed	1,632,803	30,622,564	Pokec online social network
soc-Slashdot0811	Directed	77,360	905,468	Slashdot social network from November 2008
soc-Slashdot0922	Directed	82,168	948,464	Slashdot social network from February 2009
wiki-Vote	Directed	7,115	103,689	Wikipedia who-votes-on-whom network
wiki-RfA	Directed, Signed	10,835	159,388	Wikipedia Requests for Adminship (with text)
gemsec-Deezer	Undirected	143,884	846,915	Gemsec Deezer dataset
gemsec-Facebook	Undirected	134,833	1,380,293	Gemsec Facebook dataset

Copyright © 2021 Bennett, Coleman & Co. Ltd. All rights reserved. For reprint rights

Following a Weight Loss Diet

1 Include lean protein at each meal to support your weight loss. Protein is an essential nutrient, but it's especially important for weight loss. It can help support your

Text Analysis
Journalist / Poster Analysis
Network and metadata Analysis
Original Source Analysis
Image processing and visual Analysis
Part 1
lean protein
1/4 plate
wikiHow to Lose 5 Kgs in 30 Days

FIGURE 9. Multimodal analysis of a blog.

A. PROPOSED METHODOLOGY

The credibility analysis pipeline is divided into three major modules. The first module examined the blog's platform and its contribution to the blog's overall credibility. The authors use SEO tools to analyze the platform's features and determine its overall score. Finally, this study created a regression model to calculate a blog's overall web credibility score, given a score for all its contributing features. This assists us in determining the credibility of the websites in the collected validation dataset.

Furthermore, the study examines how well-linked the platform is by running it through the well-known Page-Rank algorithm, which provided us with a score for each blog in the dataset. The second module included studying the blog text to help determine the author's credibility by inspecting the author's writing style. This section calculates a final credibility score by putting the blog through four literary tests: readability, grammar, expertise, and correctness. Finally, in the third module, the study examines the blog's images and derive an overall score of quality & authenticity.

There are five major analysis techniques employed in this study, namely, 1) Regression Analysis, 2) Web Analysis, 3) Author Writing Style Analysis, 4) Image Authenticity Analysis, and 5) Image Quality Analysis.

Each one of these five processes has different technical aspects. In the first module, the authors employ various regression models & the page rank algorithm. In the second module, they use Natural Language Processing algorithms to pre-process the material and run four tests for writing style analysis. Finally, in the third module, they develop a deep neural network using Transfer Learning to examine the authenticity of an image. BRISQUE (Blind Reference less Image Spatial Quality Evaluator) is used for Image Quality Analysis. After completing these modules, they feed the results to the X-AI model to provide correct explanations of the results. Fig (10) shows the system architecture for this project.

1) WEB CREDIBILITY

"One can identify misinformation by surveying the credibility of its website, where credibility is frequently characterized in the sense of quality and authenticity." The genuineness of web material is defined as web content credibility. It is made up of two factors: trustworthiness and knowledge. People can tell whether a webpage has both traits if it has substantial or valid content. Holding an exquisite, talented, trying webpage offers credibility to the substance. The web credibility rating problem was formulated as a prediction task, and two methods were used to model it (1) Web Analysis (Application of the Page Rank Algorithm) (2) Regression analysis.

a: WEB ANALYSIS

PageRank measures the importance of each node in the graph by counting the number of external links that connect it along

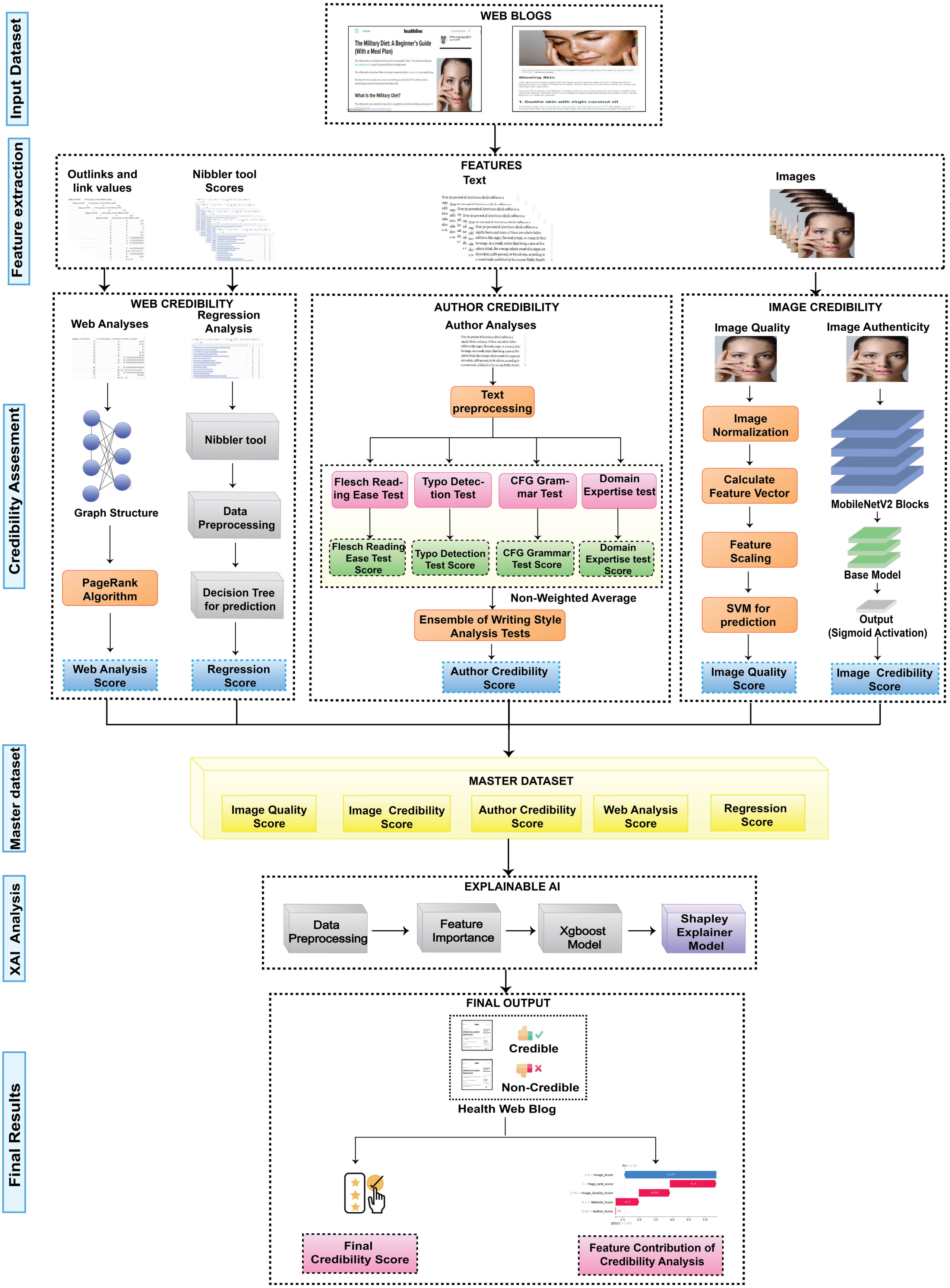


FIGURE 10. System architecture.

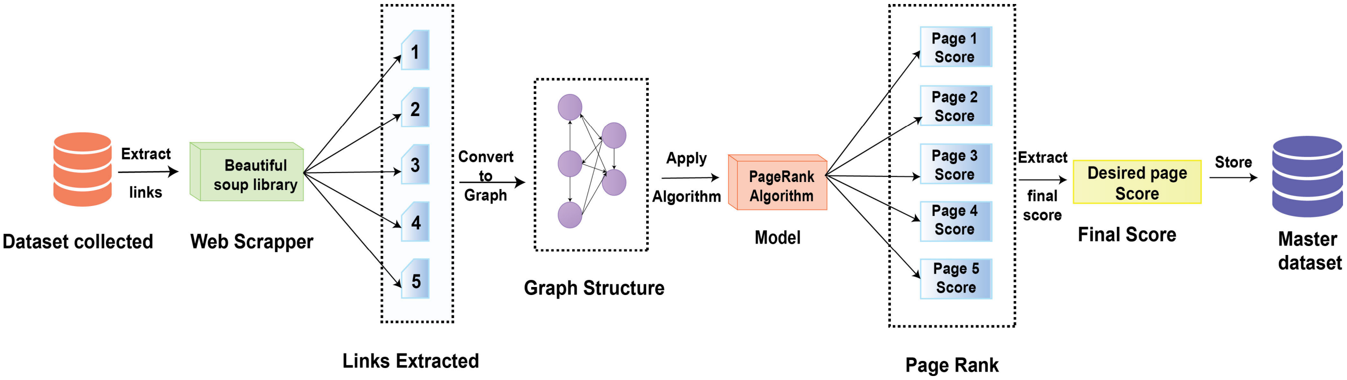


FIGURE 11. Page rank analysis.

with its source node. The underlying assumption is that a page must link to other pages to be relevant and powerful. Several segments are involved in creating this efficient model, including data collection and the development of the page rank algorithm, which are described below. Fig 11 depicts this pictographically.

i) DATA COLLECTION

The input for the ranking in the page rank algorithm is a link graph constructed using the link value of how each webpage is connected to another. To create that graph, five links for each website are extracted using the Beautifulsoup library in Python and the link value between them through their linkage is found. It comprises of three features for the input graph: the page number, out link page number, and link value (using the formula of the link matrix described below). This dataset has 1000+ web links and linkages, which, when converted to a graph and subjected to the page rank algorithm, gives the page rank of the desired web pages.

ii) APPROACH

According to Page Rank, an important site is based on its links to and from other sites. This is where the Eigen theory comes into play. The authors described the links of a page u as vectors, in which each row is either a one or a zero depending on whether a link is present to that page and then described a probability for each page. The next step is the normalization of the vector by the number of linked pages and built a link matrix L by dividing them into columns, a square matrix. The matrix L tries to represent the probability of appearing on each page. Even though the matrix was constructed from columns of outward links, the rows described inward links normalized according to their page of origin.

iii) IMPLEMENTATION

Here the page rank algorithm is built to determine the score of each web page using the method described above. Working is discussed in brief in Fig (12).

To store the rank of all web pages, the authors utilized a vector called r. For computing rank of a page u the following facts related to all its web pages must be known:

1. Rank of the webpage
2. Do they include a link to page u?
3. What is the total number of outgoing links they have?

The following expression combines these three pieces of information for webpage *u* only.

$$R_u = \sum_{j=1}^{n} L_{u,j} r_j \tag{1}$$

where the number of webpages on each website are represented with the vector r and L ∈ $R^{n \times n}$ is the link matrix relevant to page that can be shown by the formula

$$L : \left\{ \begin{matrix} \frac{1}{n_j} & if\,,\ j \in N_j \\ 0 & otherwise \end{matrix} \right\} \tag{2}$$

R_u is then equal to the sum of j = 1 to n, where n is the sum of all the webpages in the link matrix relevant to u and the j location, multiplied by the rank at the j location. This scrolls through each of the web pages, causing the rank of u to be a sum of the ranks of all the links coming from those pages, weighted by the link probability. The authors then solved the problem for all pages simultaneously by writing the same expression.

Next this expression for all the web pages is calculated by performing simple matrix multiplication.

$$r^{(i+1)} = Lr^{(i)} \tag{3}$$

Then, multiplying r by the matrix L each time gives us a new value for r. As a result, r is now an eigenvector of matrix L with an eigenvalue of one.

Eigenvalue:

Although various methods for effectively computing eigenvectors have been developed throughout time, the power technique - which multiplies a randomly picked starting guest vector by a matrix - still works effectively for solving the page rank problem. First, despite the fact that the power technique can only offer you one eigenvector because there are n for an

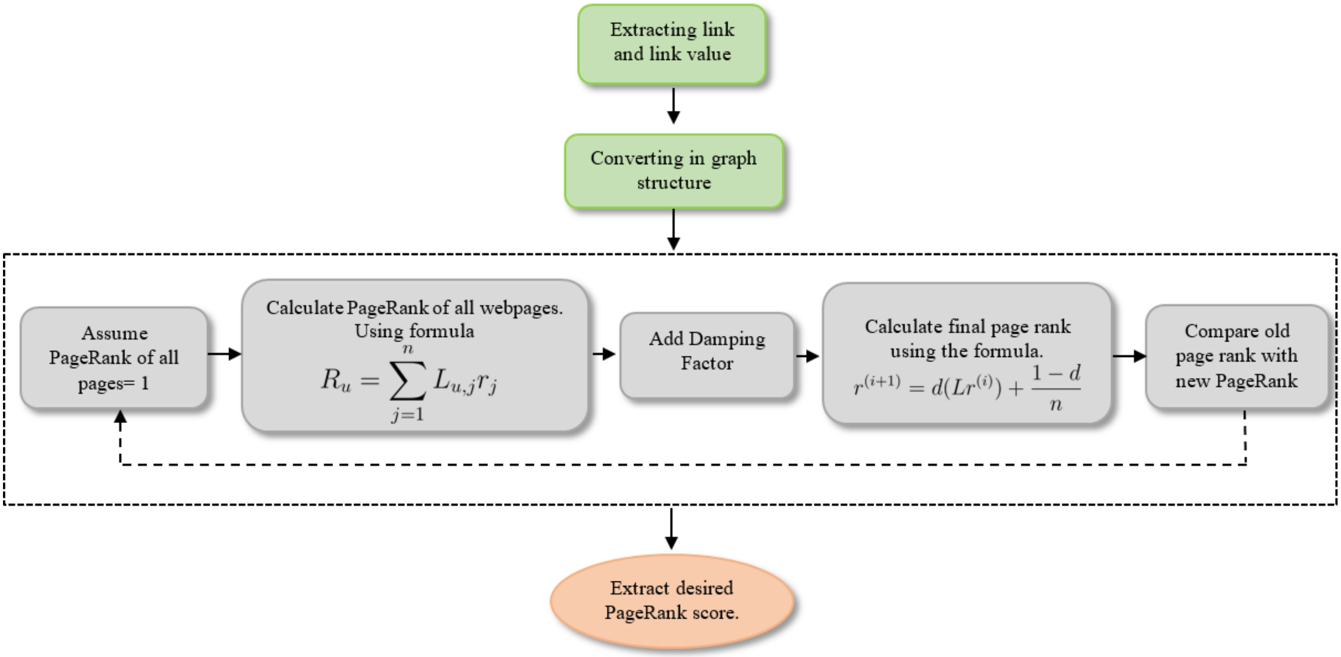

Extracting link and link value
Converting in graph structure
Assume PageRank of all pages= 1
Calculate PageRank of all webpages. Using formula
$R_u = \sum_{j=1}^{n} L_{u,j} r_j$
Add Damping Factor
Calculate final page rank using the formula.
$r^{(i+1)} = d(Lr^{(i)}) + \frac{1-d}{n}$
Compare old page rank with new PageRank
Extract desired PageRank score.

FIGURE 12. The page rank algorithm: Working.

n-page system, the only vector it gives you is the one you're looking for, with an eigenvalue of 1. Second, while this is not true for the whole webpage mini-Internet, it may conclude that practically every entry in the link matrix is zero in the real Internet, implying that most pages do not link to each other. A sparse matrix is what this is called. Multiplication may be made much easier with the help of algorithms. And then, the damping factor to calculate the final page rank can be added.

$$r^{(i+1)} = d(Lr^{(i)}) + \frac{1-d}{n} \tag{4}$$

Damping Factor:

The simple update rule can cause PageRank to collect and become stuck in certain portions of the graphs for particular graphs.

- This is fixed by assigning each node to
  - Give a d proportion of its PageRank to its neighbors (at each round).
  - Give everyone in the graph a (1-d) fraction of its PageRank.
- This also means that pages with no incoming connections gain some PageRank.
- Here damping factor is d (generally fixed to 0.85)

The impact on the actual calculation is about establishing a balance between the iterative convergence process's speed and stability. Finally, the algorithm concludes by importing the page rank scores of all the desired web pages to the master dataset.

b: REGRESSION ANALYSIS

Regression Analysis allows us to discover which factors are most important, which ones may be ignored, and which ones interact with one another. Regression analysis is used for two reasons: to predict the value of the dependent variable or to evaluate the effect of an explanatory variable on the dependent variable. So to predict the web credibility score of each web blog, we used Regression Analysis which started with data preprocessing of the dataset and then using feature extraction. We selected the most important features impacting our predictions. Using those features, we trained our dataset on 4 Regression Models, namely Linear Regression [63], Support Vector Regression [64], Decision Tree Regression [65], and Random Forest Regression [66]. These models were trained on the dataset to find the overall platform credibility score, telling us how important and credible that website is. To find that score, we evaluated our model based on three parameters, i.e., root mean squared error, R - squared error, and Explained variance error. Through these parameters, we found our best fit model through the steps described below and then predicted the overall score and stored it in master dataset. Fig (13) describes this pipeline in detail.

i) DATA PREPROCESSING

The first and most vital step in creating a machine learning model is preprocessing the raw data, making it suitable for prediction. It is required as machine learning algorithms cannot directly use data that might contain noise or has missing values. This entire process involves several steps. For the collected dataset, some of those were not necessary. The dataset and libraries necessary for the model were imported. Our cleaning procedure involved removing the 'BLOG' feature (name of the blog). The authors then checked for missing values and had a fairly clean dataset with very few missing

Pseudocode 1 Page Rank Pseudocode

```
procedureP($H$,$i$) .                                  $H$: inlink file, $i$: # of iteration
    $f \leftarrow 0.85$ .                               # damping factor: 0.85
    $o \leftarrow H$   .                                # outlink of H
    $i \leftarrow H$   .                                # inlink of H
    $m \leftarrow H$   .                                #no of pages from H
    for all $pr$ in the graph do
        $op[pr] \leftarrow \frac{1}{3}$                 #start the Pagerank
    end for
    while $i > 0$ do
        $dp \leftarrow 0$
        for all $pr$ that has no out-links do
            $dp \leftarrow dp + f * \frac{op[ip]}{o[ip]}$    #obtain Pagerank without out-links from pages
        end for
        for all $pr$ in the graph do
            $np[pr] \leftarrow dp + \frac{1-f}{m}$       # jump obtain Pagerank from random
            for all $ip$ in $i[pr]$ do
                $np[pr] \leftarrow np[pr] + \frac{f*op[ip]}{o[ip]}$    #with inlinks obtain Pagerank
            end for
        end for
        $op \leftarrow np$                              #reform Pagerank
        $i \leftarrow i - 1$
    end while
end procedure
```

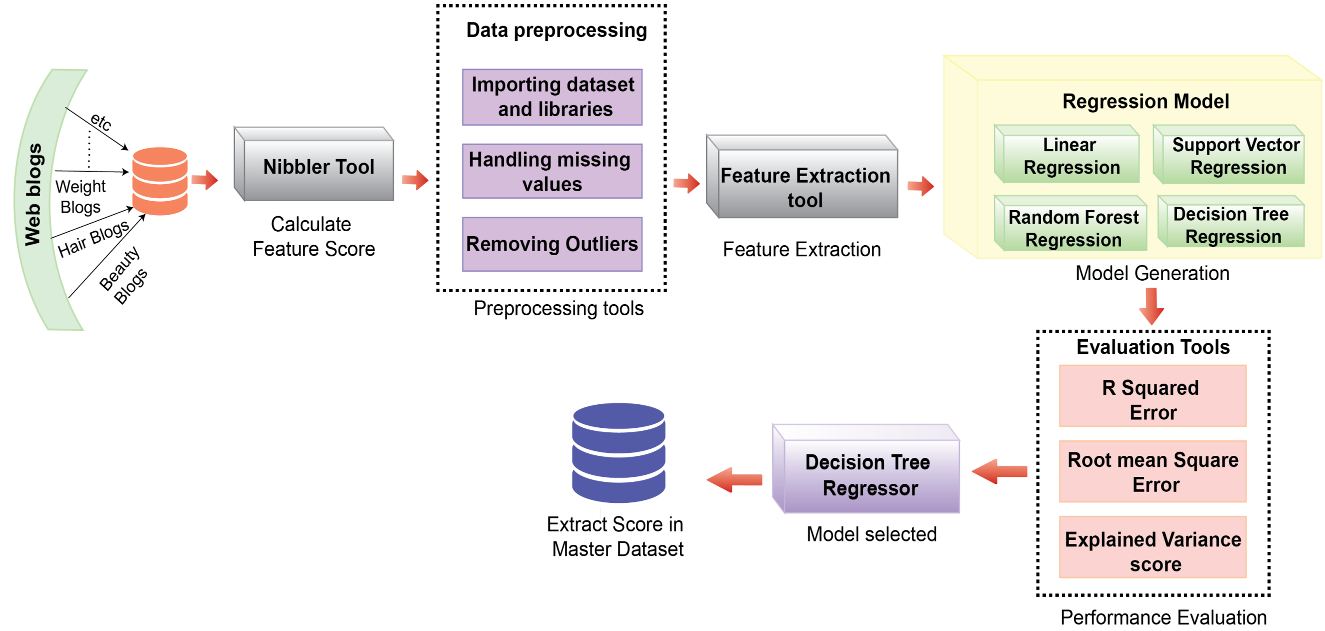


FIGURE 13. Regression analysis.

values. It was found out that Freshness and Twitter columns had few missing values that were handled using the Imputer class of sklearn preprocessing library.

ii) FEATURE EXTRACTION

When it comes to feature engineering, choosing the most important subset of features and removing the features that have the least impact on performance, and achieving optimal performance for a given ML task are critical. An efficient set of feature subsets is the most important factor in developing the machine learning model, which reduces the likelihood that every model develops an overfitting problem. Using Correlation, 11 out of 13 features were found necessary for prediction.

iii) BUILDING THE MACHINE-LEARNING ALGORITHM

This section represents the baseline algorithms used for regression analysis, namely Linear regression, Support. Vector regression, Decision tree regression, and Random forest regression.

Linear Regression:

In Linear Regression, a linear relationship is formed between two or more features of the dataset, producing an outcome of the dependent variable. This function performs a regression procedure. The regression procedure produces a target prediction value based on independent variables.

As the dataset has more than one independent variable, the study applied multiple linear regression, which works on assumptions like - target and predictor variables must have linear relationships. Additionally, MLR assumes no or little multicollinearity (correlation between independent variables).

Following these assumptions, the predictions were made based on the formula below:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots + \beta_p x_{ip} \in \tag{5}$$

where:

1. The dependent or anticipated variable is yi.
2. The y-intercept(β_0), or the value of y when both xi and x2 are zero, is 0.
3. The regression coefficients β_1 and β_2 show the change in y as a function of a one-unit change in xi1 and xi2.
4. For each independent variable, β_p is the slope coefficient.
5. The random error is ϵ

Support Vector Regression:

The supervised learning algorithm Support Vector Regression is used to predict discrete values. In SVM, a hyperplane is a straight line that fits the data. The idea behind the support vector machine approach is to locate a reference point that defines a hyperplane over n-dimensional space that categorizes data points. The Support Vectors run the length of the hyperplane, assisting in its positioning and orientation. The SVR, unlike other regression models, attempts to fit the best line within a given threshold value. The distance between the hyperplane and the boundary line is the threshold value. Thus, any hyperplane that satisfies the SVR should satisfy:

$$-a < y - wx + b < +a \tag{6}$$

where y =wx+b is the equation of hyperplane. Only those within the decision border and have the lowest error rate, or those within the Margin of Tolerance, are used. This results in a more accurate model.

Decision Tree Regression [62]:

Decision trees provide models of classification and regression in the form of a tree structure. A dataset is segmented into smaller and smaller subsets as a decision tree is developed incrementally. Finally, a tree containing a decision node and leaf nodes is built. Each decision node represents the property being tested for. Each leaf node reflects a numerical target decision. The root node is the topmost decision node in this tree, and it correlates to the best score.

Multiple techniques are employed in Decision Trees to split a node into two or more sub-nodes. The study utilized the ID3 algorithm to generate the model. Using a top-down greedy search with no backtracking, this approach constructs decision trees from the space of possible branches. Iterates over the very unused attribute of the set S, calculating Entropy (H) and Information Gain (IG) of this property, using the original set S as the root node.

The method then divides the set into subsets based on the attribute with the lowest Entropy or highest Information Gain. It keeps repeating recursion on subsets, selecting only attributes that have never been placed before. The study used entropy criteria to determine whether attributes should be placed at the root or as internal nodes within each node (The entropy of information being processed represents its randomness). A branch with an entropy of zero is a leaf node, and a branch with an entropy greater than zero requires additional splitting, according to the ID3 method. It used the following mathematical function for multiple attribute entropy: -

$$E(T, X) = \sum_{c \in X} P(c)E(c) \tag{7}$$

where T→ Current State and X→ Selected attribute

Based on each attribute's value, the attributes are sorted as follows: the highest value is placed at the root.

Random Forest Regression:

To tackle regression problems, a Random Forest [63] is an ensemble approach that uses numerous decision trees and a technique known as Bootstrap and Aggregation, sometimes known as bagging. This method aims to integrate decision trees instead of depending on individual trees to produce the final result. Numerous decision trees can be merged into distinct basic learning models using Random Forest. The authors create a sample dataset for each model by randomly selecting rows and attributes from the training dataset. Multiple decision trees are used to create RF, which may yield in more accurate and reliable outcomes. The ultimate output in a regression problem is the average of all the outputs.

iv) PERFORMANCE EVALUATION METRICS

A regression model's skill or performance must be measured as an error in predictions. After training with the above machine learning models, the study evaluated their performance using 3 evaluation metrics:

a. *R- Squared error* - Using a linear model and R-squared formula, an R-squared statistic measures how much of the variation in a response variable is explained.

$$\mathrm{R}_2 = \frac{\text{Variance Explained By the Model}}{\text{Total Variance}} \tag{8}$$

It is always between 0 and 100% where:

i. 0% reflects the fact that the model is unable to explain any response variability around its mean.
ii. 100 % indicates that the model is fully statistically sound because it explains all the variation of the response data around the mean.

The R-squared value is a measure of how well a model fits the data. The greater the value, the better.

b. *Root Mean Squared Error* - The Root Mean Square Error (RMSE) measures the deviation of residuals. from the predictions (prediction errors). This study tells us how condensed the data is along the line of best fit. It can range from 0 to ∞. Lower values of RMSE indicate a better fit. It uses the following formula to find the error:

$$RMSE_{fo} = \left[\sum_{i=1}^{n} (z_{fi} - z_{oi})^2 / N \right]^2 \tag{9}$$

where:

i. $\Sigma =$ summation(“add up”)
ii. $(z_{fi}–z_{oi})^2 =$ differences, squared
iii. $N =$ sample size

c. *Explained Variance* - Explained variance is used to measure the difference between a model and actual data. More specifically, it is the part of model total variance which cannot be attributed to error variance. It works on the following formula:

$$EV(y, y_i) = 1 - \frac{Var(y - y_i)}{Var(y)} \tag{10}$$

Var (y – yi) and Var(y) is the variance of prediction errors and actual values, respectively; scores close to 1.0 are highly desired, indicating better squares of standard deviations of errors. Having a larger percentage of the variance explained suggests a stronger degree of association. Also, it indicates better prediction.

The study aims to achieve a model with a high R squared error, high variance score, and low root mean square error.

2) AUTHOR CREDIBILITY

Many people rely on blogs as their primary source of information. In this case, the author of the blog bears the responsibility of disseminating credible information. The most effective way for an author to communicate with his or her readers is through the blog’s content. The author can express his thoughts, opinions, and views on various topics using this medium. However, this communication becomes more rigid in the case of health blogs. Every remedy, tip, and fact shared by the writer must be credible, as gullible readers may end up experimenting on their bodies. In some cases, this may have long-term consequences. To tackle this issue, the authors propose a thorough analysis of the content of the blog.

a: PRE-PROCESSING

After scraping the content of all the blogs in the dataset, the authors subject it to pre-processing to prepare it for analysis. Since this text is in its most raw form, the authors start with some basic data cleaning techniques. Next all potential sources of noise were removed from the data. This includes removing any HTML tags, punctuation, or special characters that do not contribute to the text’s meaning. The text is also made free of all abbreviations, white spaces, numbers, and other diatrics that it may contain. This is then followed up with the removal of stop words. Stop words are commonly used words such as “a,” “an,” and “the” that add no value to the sentence. A typical search engine is designed to ignore such terms. The next layer of pre-processing involves Lemmatization. Lemmatization tries to diminish inflectional structures to a solitary base structure. Unlike stemming, it does not simply eliminate intonations. Instead, it consults lexical knowledge bases to determine the proper base forms of words. At the end of this step, the words are now in their root base structure. After completing all these steps, the data is ready for analysis.

b: IMPLEMENTATION

The most distinctive feature of an author’s blog tends to be his/her writing style. While a judgment on the writing style is highly subjective, an analysis can help derive multiple inferences. The writer’s sentiments, grammar knowledge, domain expertise can all be figured out via analyzing how they communicate information. While verifying the claims/ facts they state are out of scope for the project, the authors focus on analyzing the various aspects of their composition. The authors perform four tests, namely a Readability, Grammar, Typos & Domain expertise test. Fig (14) describes the pipeline in detail.

i) COMPUTING THE READABILITY SCORE

Readability is defined as the ease with which a reader can perceive & understand a piece of text. An average adult reads at a 7th to 9th Grade level. To make a blog highly accessible to readers, the threshold of readability needs to be maintained. A health blog should be simple & concise for its readers to understand while maintaining a technical tone. Thus, a blog has to be not only accessible but also ethnically appropriate. For an author, it becomes vital to recognize the readability level they should maintain to suit the expectations of their readers. Once they establish this level, they must maintain this consistency in their writing. Blogs where authors tend to practice this are more preferred by all kinds of readers. To determine the readability score for a piece of writing, the authors applied the Flesch Reading Ease test [64], [65]. Dr. Robert Flesch developed the Flesch formulae for readability after observing that most languages are designed to make reading as difficult as possible. He argues that most sentences were lengthy & words esoteric, making simple meanings more complicated.

To implement the test, the authors calculate the total number of words, sentences & syllables in the text. Syllables are the basic single units of speech. To calculate the total number of syllables, the following approach is used.

```
For all word in text do:
For all vowels in word:
        syllable+ =1
  If word endswith [‘es’,’ed’,’e’]:
        syllable−=1
  If word endswith [‘le’]:
      syllable+ =1
```

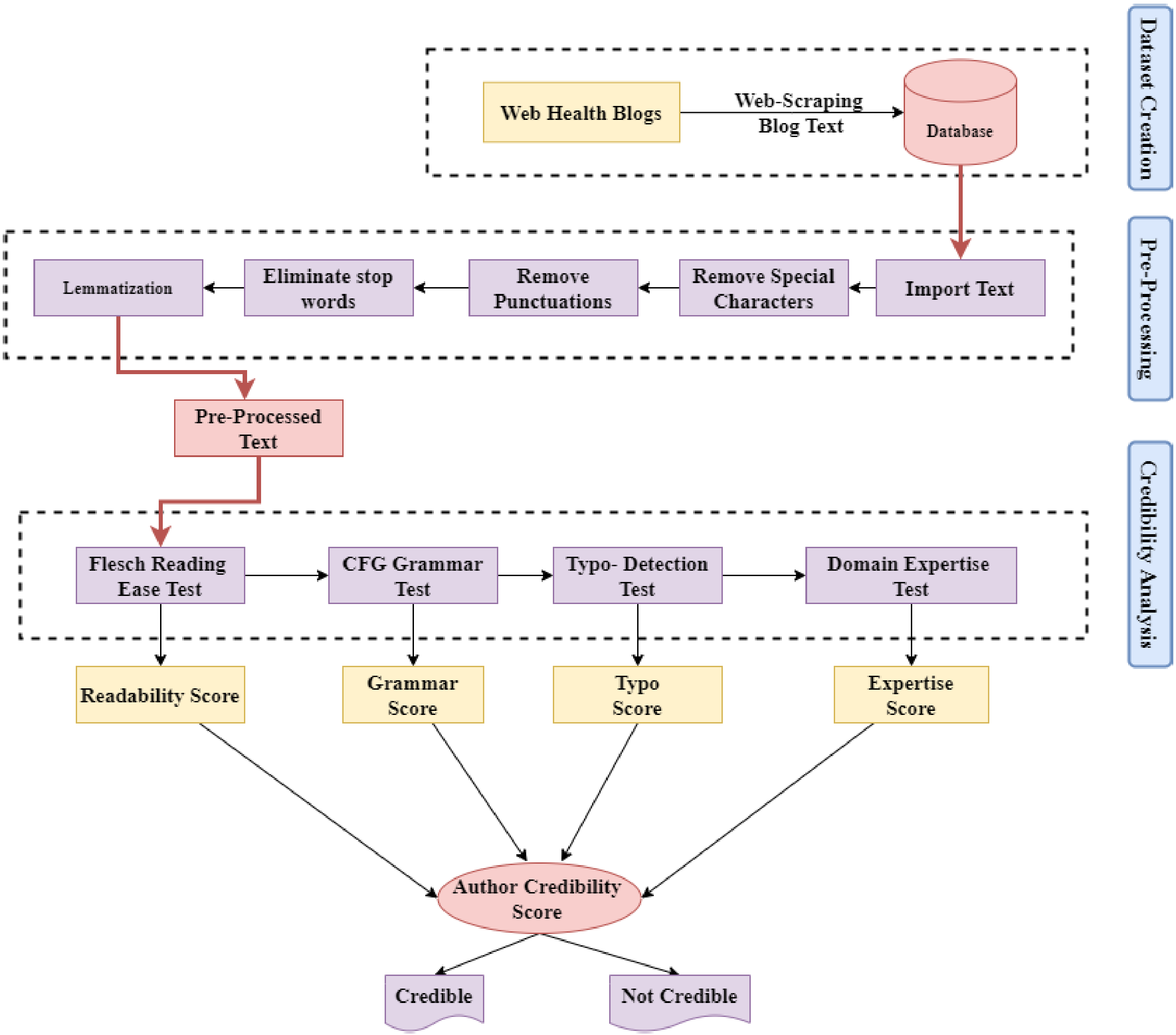


FIGURE 14. Author credibility analysis pipeline.

Once all the necessary information has been gathered, the Flesch Reading Ease Test is administered as follows:

$$\begin{aligned} &Readability\_score \\ &\quad = 206.835 - 1.015 * (no\_of\_words/no\_of\_sentences) \\ &\qquad - 84.6^{*}(syllable/no\_of\_words) \end{aligned} \tag{11}$$

Where: no_of_words: Total Number of Words in corpus
No_of_sentences: Total number of Sentences in Corpus.
Syllable: Total number of Syllables

The readability score obtained from this test lies in a range of 0-100. The range is scaled down to 0-10 for easier future calculations. The higher the readability score, the easier the piece of text is to read. For health blogs, a readability score of 6-7 is highly preferable. Because health blogs cater to a broader audience, they must balance readability and technical terms.

ii) COMPUTING THE GRAMMAR SCORE

Aside from readability, the majority of popular blogs have good grammar. A blog that follows all grammar rules is more likely to aid the reader's comprehension than one that contains grammatical errors. For this study, the authors have considered only English blogs. English is a context-free language, which means it adheres to Context-Free Grammar rules [66]. The production rules of formal grammar are said to be "context-free" if they can be applied regardless of the context of a nonterminal. The authors examine the grammar of all the blogs in the dataset by checking the structure of all sentences, converting them into parse trees, and determining whether they follow the Context Free Grammar's production rules [67] To generate this CFG for the sentences in the text, the authors used the spacy library. Spacy parsed and generated dependency graphs. The trees produced by a dependency grammar can have a one-to-one mapping to the trees produced by a context-free grammar. Spacy, on the other hand, does not use explicit grammar to parse. Rather, a neural network is used to determine how to place the dependency relationships on a case-by-case basis. According to a specific grammar, the neural networks that power the parser and tokenizer were trained on corpora that had been hand annotated. The grammar is evaluated on a scale of 0 to 10, counting the number of sentences that follow the production rules and are deduced to parse trees from the total number of sentences in the blog. This can be formulated as follows:

$$GrammarScore = (Grammatically\ right\ sentences \\ /Total\ number\ of\ sentences) * 10 \quad (12)$$

A score of 10 indicates that the author's grammar is most appropriate, as each sentence follows the proper structure for the English language. A grammar score of 8+ is ideal for a credible health blog.

iii) COMPUTING DETECTED TYPOS SCORE

Typos, short for typographical errors, are usually mistaken in the spellings of words that more than often are misprints or typing mistakes. While typos can be overlooked in most cases, it is preferable for a good and credible health blog if it is completely free of them. Their presence not only makes the author appear untrustworthy, but it may also result in a lack of clarity and unnecessary confusion for the readers that could have been avoided. Thus, for a blog to be trustworthy, the content it shares must be free of such minor errors. To check this, the authors ran a typo detection test on the blogs in the dataset. The authors use the TextBlob library to correct the mistakes in the original text. The corrected text is compared with the original text to determine the number of misspelled words by the author. This allows us to calculate the overall spell check score for the text.

TextBlob is a Python package that helps users prepare text-based data. This module provides a stable API for common normal language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and so forth. TextBlob's built-in correct() function takes a corpus of text and returns a repaired version with minimum spelling problems. A simple implementation of the same is as follows: For each blog text in the dataset, do:

Text_blob = TextBlob(blog_text) # create Text blob instance

Text_blob. correct() # Use the correct() function for spell check.

While the correct method can help bring typo rates down significantly, some mistakes persist. This is because of a phenomenon called "overcorrection." Sometimes, the function does not have sufficient information regarding the context of a word and thus may correct a word already spelled correctly. However, there is no perfect spelling corrector since most languages are highly contextual; thus, the small overcorrection can be ignored. TextBlob provides results that are quite suitable for an average user. After generating the corrected text, the authors compute its semantic similarity with the original text. This similarity score is evaluated on a range of 0-10. A higher similarity score would mean the original text was already free of typos; on the other hand, a lower similarity score would suggest disparities in the original text. With the typo detection test score now computed, the authors complete inspecting yet another aspect of an author's writing style.

iv) COMPUTING DOMAIN EXPERTISE SCORE

After reviewing a blog's readability, grammar, and spelling, the authors move on to determining its technical relevance. It is critical for a health blog to provide valid explanations/medical evidence for any remedy or tip suggested by the author. As a result, an author who uses proper medical terms has more domain knowledge than one who writes the article in the most generic terms. Verifying the author's claim/medical evidence may be a future scope of this project, but for now, the research only determines the amount of technical jargon used by the author. To identify these medical terms, the study uses a glossary of all medical jargon. Harvard Medical School offers a public dictionary of health-related terms. This dictionary contains over 5000 terms covering every letter of the alphabet. The authors examine each blog in the database for the presence of these terms. The percentage that these words cover in a blog text help determine the domain expertise score. Each score is scaled down to the range of 0-10. A higher score would suggest that the author has used the most appropriate technical terms & the blog is more scientifically apt.

At the end of these tests, there are 4 scores, namely, 1) Grammar Score, 2) Typo Detection Score, 3) Readability Score & 4) Domain Expertise Score. Each of these scores contributes equally to determining the author's writing style. As a result, a non-weighted average of each of these scores yields the overall author credibility score.

3) IMAGE CREDIBILITY

Humans are visual learners; when concepts or ideas are visualized, they tend to grasp them easily. Images are a great tool for authors to use when communicating ideas to their readers. The entire reading experience becomes more immersive, with a proper balance of texts and images in the blog. Today, more than 657 billion images are uploaded on the internet every year. It becomes critical to be able to determine the authenticity of an image. There is a plethora of images on health blogs that depict the effects of a particular remedy or product on the human body. While these images can help readers gain clarity by visualizing the results, they can be dangerous if they are doctored. Some authors use photoshop to demonstrate that their remedy worked perfectly. An unsuspecting user falls prey to their deception and experiments with the remedy or

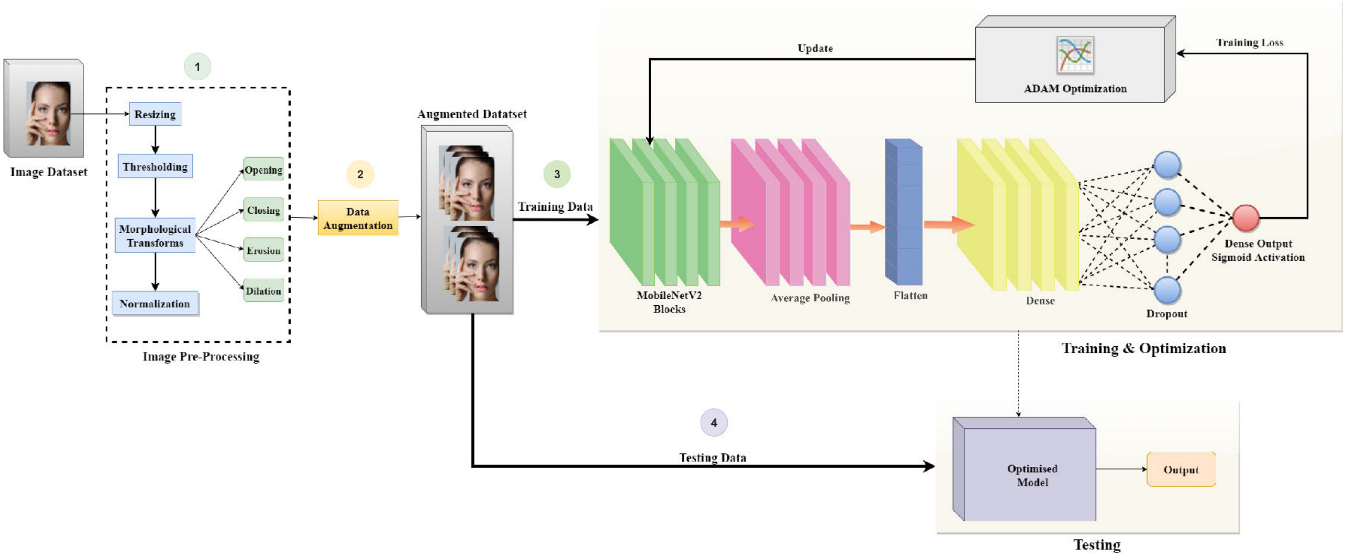


FIGURE 15. Image credibility analysis.

product on their bodies. In some cases, this can have a serious impact on the reader's health. Examining the credibility of a blog's images is thus equally important when determining its overall credibility.

A good health blog has not only genuine images but also maintains high-quality images. When images on a blog are distorted or in the wrong dimensions, they have a negative impact on the readers. As a result, while assessing the fakeness quotient of images, we must also consider their quality. Our research calculates the image credibility score for a blog by 1) determining how fake the images are and 2) determining the quality of these images. A blog with images with a low fakeness quotient and high quality has a higher image credibility score.

a: VERIFYING THE AUTHENTICITY OF IMAGES

Fake/Photoshopped Images have distinctive features that set them apart from real images. For example, the "after" images in blogs highlighting fake weight or color tone lightening can be easily set apart from their before counterparts. The average user easily makes these distinctions, but a few may miss them. A deep learning model is trained that recognizes and identifies the features that distinguish a fake image from its real counterpart to automate the entire process. Next, transfer learning techniques are used to develop a high-performing network, having trained it with over 800 images collected from all the blogs in the dataset. Fig (15) describes the image authenticity analysis pipeline.

i) DATA PRE-PROCESSING

Before feeding out images to the neural network, the study performs pre-processing [68]. Having stored the images in two separate directories (Real & Fake), the algorithm iterate over all images in the dataset, set their corresponding labels & subject them to pre-processing. This step starts by resizing all the images in the dataset to have a standard size. Since most images are of different types & colors, it is vital to determine a standard channel and size. For the dataset, the algorithm resizes the images to have a size of (224,224) with standard RGB encoding. Having done this, the authors move on to performing morphological transformations on the images. Morphological transformation refers to changing the shapes & forms of images to maintain a uniform structure for analysis. In this case, the authors subjected the images to various types of transformations. Next, Thresholding is used to convert each image to its corresponding binary form by setting a threshold for Pixel values. This is followed up by four processes: Opening, Closing, Erosion & Dilation. Erosion reduces the size of bright areas while increasing the size of dark areas. On the other hand, Dilation has the exact opposite effect: it shrinks dark regions while enlarging bright regions. The opening can be used to remove small bright spots as well as connect small dark cracks. This has the effect of "opening" up (dark) gaps between (bright) features. It is usually the process of erosion followed by dilation. On the other hand, closing is dilation followed by erosion that can help reduce the small dark spots. The final step of the preprocessing was normalization. This refers to rescaling the pixel values to fit within a specific range. One of the reasons for doing so is to assist with the problem of propagating gradients. After completing these steps, the image array is now ready to be inputted into the neural network.

ii) DATA AUGMENTATION

The more data is feed into the neural network, the better its performance. Data Augmentation [69], [70] refers to this technique of increasing the quantity & variance of the original data by performing certain transformations on

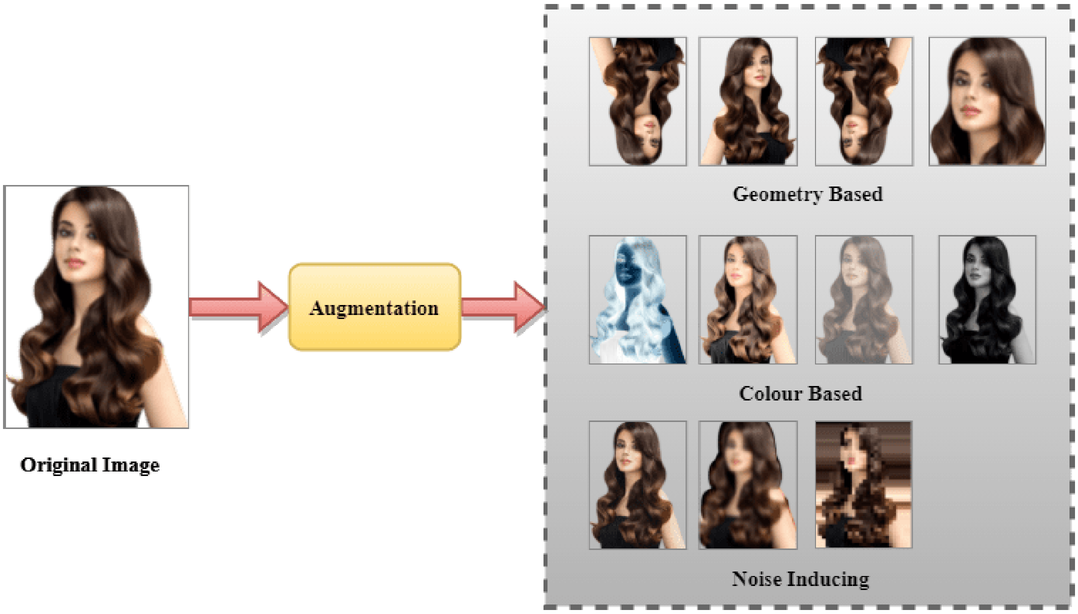


FIGURE 16. Data augmentation techniques.

it. This work implemented data augmentation methods to subject the model to a wider range of images to prepare it to make better predictions. Methods like zooming the image, flipping the image by a certain degree, adding noise to it, blurring the image, experimenting with its brightness (illumination/dullness), image translation, etc., are used. Each technique is applied to every image in the dataset, thus exposing the model to a much larger dataset [71], [72]. Fig (16) shows the various data augmentation techniques employed.

b: BUILDING THE MODEL

The authors used transfer learning [73]–[75] to build the classification network. The authors chose the MobileNetV2 [76] architecture as the pretrained head model, which is customized by adding a few layers on top for classification. MobileNet V2 is a model created by Google. It utilizes depth-wise separable convolution as a means of improving efficiency over its predecessor, MobileNetV1.But V2 adds two new architectural features: a) linear bottlenecks between layers and b) shortcuts between bottlenecks. In the bottleneck layer, two residual connections follow an inverted structure. The intermediate expansion layer filters are based on lightweight depth-wise convolutions, which provide non-linearity. In MobileNetV2, a full convolution layer with 32 filters follows a residual bottleneck layer with 19 filters. MobileNetV2 models are faster while maintaining the same accuracy across the entire latency spectrum. There are 2x fewer operations and 30% fewer parameters in the new models which has been trained on the ImageNet dataset, comprised of 1.4 million images classified into 1,000 classes, has pre-trained the system. They include top parameter of this model is set to be false so as to not include the classification layers. Then the model is fine-tuned by adding to it a few Pooling, Dropout & Dense layers. Our final Dense layer has a single output & is activated by the sigmoid activation function. The sigmoid function scales the output to lie within a range of 0 to 1. Since this is a binary classification problem, an output of >0.5 will classify the prediction as "TRUE" or otherwise as "FALSE".

Morphological transforms were used to sharpen the images in the dataset. Most images scraped from this websites were compromised in terms of quality & visibility. Morphological transforms like dilation (that added pixels to boundaries) & erosion (that removed pixels from boundaries) work well on colour images and help improve brightness, remove small anomalies & also fills holes and broken areas. MobilenetV2 is lightweight and at the same time does not compromise on accuracy. Since the future scope of this study extends to creating a browser extension/mobile application for credibility analysis, it was more convenient to retrain this model for use in the browser or mobile.

c: TRAINING THE MODEL

The fine-tuned model is trained using two approaches. 1) With Data Augmentation 2) Without Data Augmentation. This is done to compare the results generated by the same model, trained with the different quantum of data. The model is trained on the model to run over 100 epochs with Adam optimizer. An 80:20 split is done on the dataset, with the training images having 80% of the dataset. The model is fed these training images & its predictions are checked on the remaining 20% validation images of the dataset.

d: PERFORMANCE EVALUATION METRICS

To evaluate how well the model did with and without data augmentation, the authors measure its performance against a

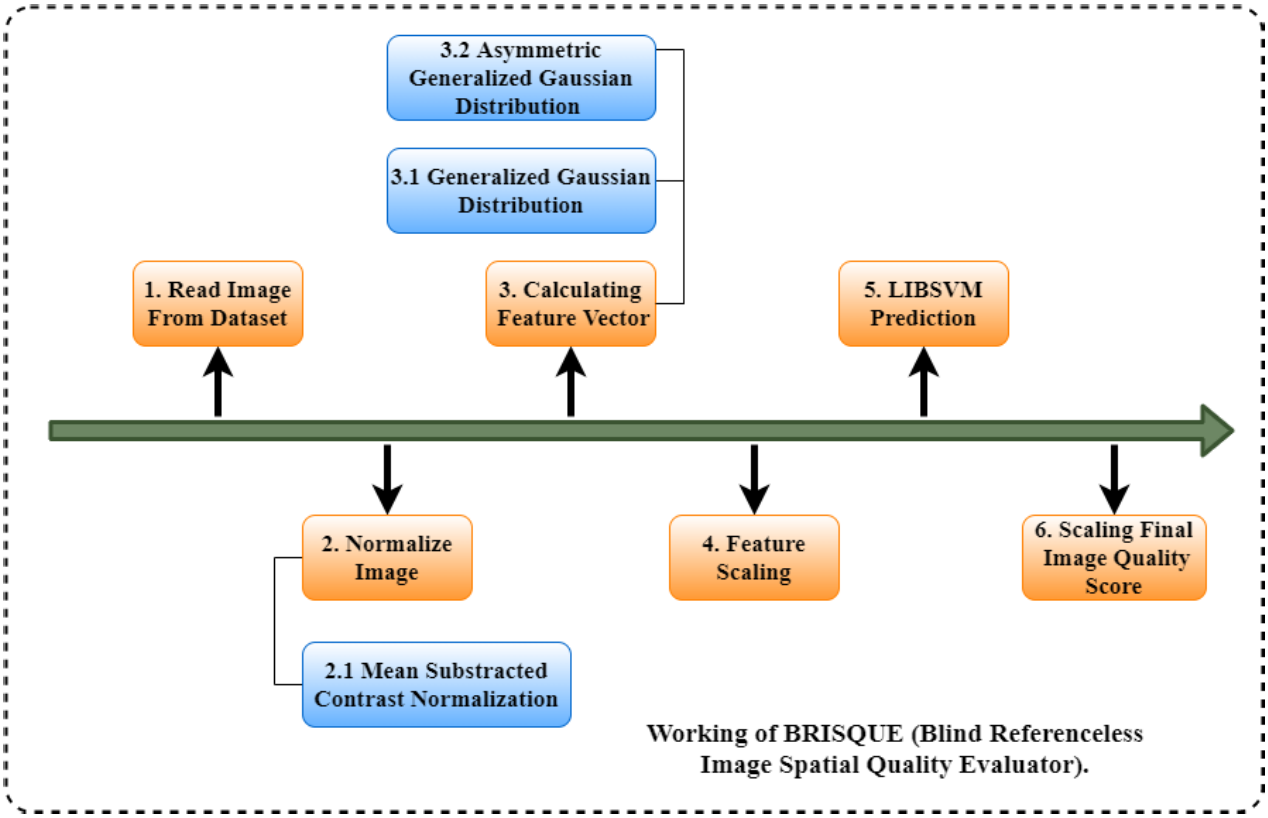

3.2 Asymmetric Generalized Gaussian Distribution
3.1 Generalized Gaussian Distribution
1. Read Image From Dataset
3. Calculating Feature Vector
5. LIBSVM Prediction
2. Normalize Image
2.1 Mean Substracted Contrast Normalization
4. Feature Scaling
6. Scaling Final Image Quality Score
Working of BRISQUE (Blind Referenceless Image Spatial Quality Evaluator).

FIGURE 17. Image quality analysis pipeline.

few metrics. Since the collected dataset was almost balanced, it can also rely on accuracy. Other than this, the F1-score, recall, precision, confusion matrix & ROC-AUC was also used. Curve to measure how well the model has made predictions. This step starts by creating a confusion matrix for the predictions. The confusion matrix helps enlist the true positives, true negatives, false positives & false negatives for the dataset. This provides additional insights into the model's performance and the types of errors the model has made the most or least. This is followed up by calculating the precision & recall. Precision measures the number of correct positive predictions made, while recall measures the correct number of positive predictions made from all the positive predictions. The F1 score combines precision and recalls into a single measure that captures both properties. Finally, the ROC curve is plotted to visualize the probability of prediction of the outcomes. The false-positive rate is plotted against the true positive rate for several candidate threshold values between 0 & 1. All these metrics help make improvements to the model to enhance its performance.

Our neural network thus helps us determine the credibility/genuineness of the images in the blog. Each image is given a credibility score in the range of 0-10. A higher score would mean that the image does not have any detectable doctored features and is quite genuine. On the other hand, a lower score would mean that the image is highly likely to have been photoshopped or modified somehow.

e: CALCULATING THE IMAGE QUALITY SCORE

By examining the pixel data, an image can be classified as noisy or blurry. However, in many cases, other aspects of image quality are impossible to be examined. Image Quality Assessment has many different techniques. In this study, the authors make use of BRISQUE (Blind Reference less Image Spatial Quality Evaluator) [77]–[79]. This falls under the category of No Reference IQA metrics. As the same suggests, for No Reference IQA, the study does not provide the algorithm with any reference to compare the input image against. The authors make use of the OpenCV library in Python to help carry out this task. Fig (17) describes this pipeline in detail.

f: PRE-PROCESSING: NORMALIZATION

This step initiates by subjecting each image in the dataset to normalization. After normalization, the pixel intensities of a good quality image follow a normal distribution, while those of a distorted image do not. By calculating how much an image differs from its normal distribution, the study can calculate its distortion. This study uses the method of normalization called the Mean Subtracted Contrast Normalization (MSCN). To calculate the coefficients, the following formula is applied:

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{13}$$

where,

$I(i,j)$: Image Intensity for a pixel at position (I,j)

$\hat{I}(i,j)$: Luminance

$\mu(i,j)$: Local Mean Field

$\sigma(i,j)$: Local Variance Field.

This can be calculate as :

$$\sigma = \sqrt{W * (I - \mu)^2}$$
$$\mu = W * I \tag{14}$$

where W is the Gaussian Blur Window function.

Images of natural and distorted types can be distinguished not only by their pixel intensity distributions, but also by their relationship to their neighbors. Pairwise products of an MSCN image with a shifted version of the MSCN image to capture neighborhood relationships. Thus, this pairwise product is found as:

$$\begin{aligned} H(I,j) &= \hat{I}(i,j)\,\hat{I}(i,j+1)\\ V(I,j) &= \hat{I}(i,j)\,\hat{I}(i+1,j)\\ D1(i,j) &= \hat{I}(i,j)\,\hat{I}(i+1,j+1)\\ D2(i,j) &= \hat{I}(i,j)\,\hat{I}(i+1,j-1) \end{aligned} \tag{15}$$

After deriving these images of original size in different orientations, the study calculates a feature vector of fixed size 36*1. Fitting the MSCN image to Generalized Gaussian Distribution yields the first two elements of the feature vector. Generalized Gaussian Distribution covers a large family of probability distributions. The main aim of this distribution is to attach a shape parameter to an otherwise normal distribution of pixels. Then, an Asymmetric Generalized Gaussian Distribution (AGGD) where the points occur at different or irregular frequencies is fitted for each of the four pairwise product images. It not only portrays the data accurately with various statistical distributions but also involves asymmetry. We then estimate the parameters of AGGD which is necessary to allow the distribution to better fit the data. As a result, the feature vector now has 18 elements. The image is reduced to half its original size, and the process is repeated to generate 18 new numbers, bringing the total to 36.

g: GENERATING THE IMAGE QUALITY SCORE

After scaling each of these feature vectors between a range of -1 to 1, they are fed to the model to predict the quality score. Traditionally, we can create the own model trained on a dataset to make predictions; however, in the study, the authors use LIBSVM. LIBSVM is a library that provides support for Classification (Binary & Multiclass) using the support vector machines (SVM) classifier. By loading the trained model first and then predicting the probability with the model's support vectors, LIBSVM is utilized to predict the final quality score. The score is scaled to lie in a range of 0-10. An image with a higher score has greater quality & thus, in turn, makes the blog more credible.

V. EXPLAINABLE AI

Understanding why a model made a particular choice is critical in any fake news detection situation. It reveals why the content was deemed fraudulent and provides fact-checkers with the information that most influenced the conclusion [80], [81]. A common method for explaining model decisions is to calculate the impact of each attribute on the decision. The increase in the model prediction error when a feature's value is permuted, which dissociates the feature from the result, can be used to calculate feature relevance [82] Permuting the values of a feature raises the correct conclusion.

On the other hand, the feature is insignificant if it maintains the model error constant because it ignores the feature when making a decision. The Shapley values are used in this study to produce a fair division scheme that specifies the elements that must be taken into account while assigning the overall importance among the attributes. Fig (18) shows the stages of AI Explainability.

A. DATA COLLECTION

To check the feature contributions, the study used the scores present in the master dataset that are gathered from the Multimodal Credibility Analysis pipeline scaled to a range of 1-10. This master dataset includes the ***website_score*** and ***page_rank_score*** from web credibility analysis, ***Author_score*** from author credibility analysis and ***image_quality_score,*** and ***image_score*** from Image credibility analysis. The authors determine the overall credibility score of each blog by taking a non-weighted average of each of these scores. With a threshold value of 6 for credibility, the study classifies each blog as credible or not. This dataset is fed to Explainable AI techniques to give a detailed explanation of the contributions of each of these scores.

B. MODEL GENERATION

The authors examined various features that could have non-linear interactions; therefore, capturing the impact of these interactions demands a highly flexible classification system. As a result, the study puts different classification methods to the test, including K-nearest neighbors, Random Forest, Decision Tree, and XGboost [83]. The time complexity largely depends on the procedure for selecting the best attribute to split and the split point. The two parameters that played a key role in the analysis:

- number of attributes;
- number of training examples

The expensive part was computing the best split point for continuous attribute (this is essentially discretization), and selection the best attributes from among the set of candidate attributes to split on.

So the time complexity is quadratic in the number of attributes *(denoted a) and linear in the number of websites (denoted n), that is, O(n * a^2). Space complexity of an algorithm denotes the total space used or needed by the algorithm

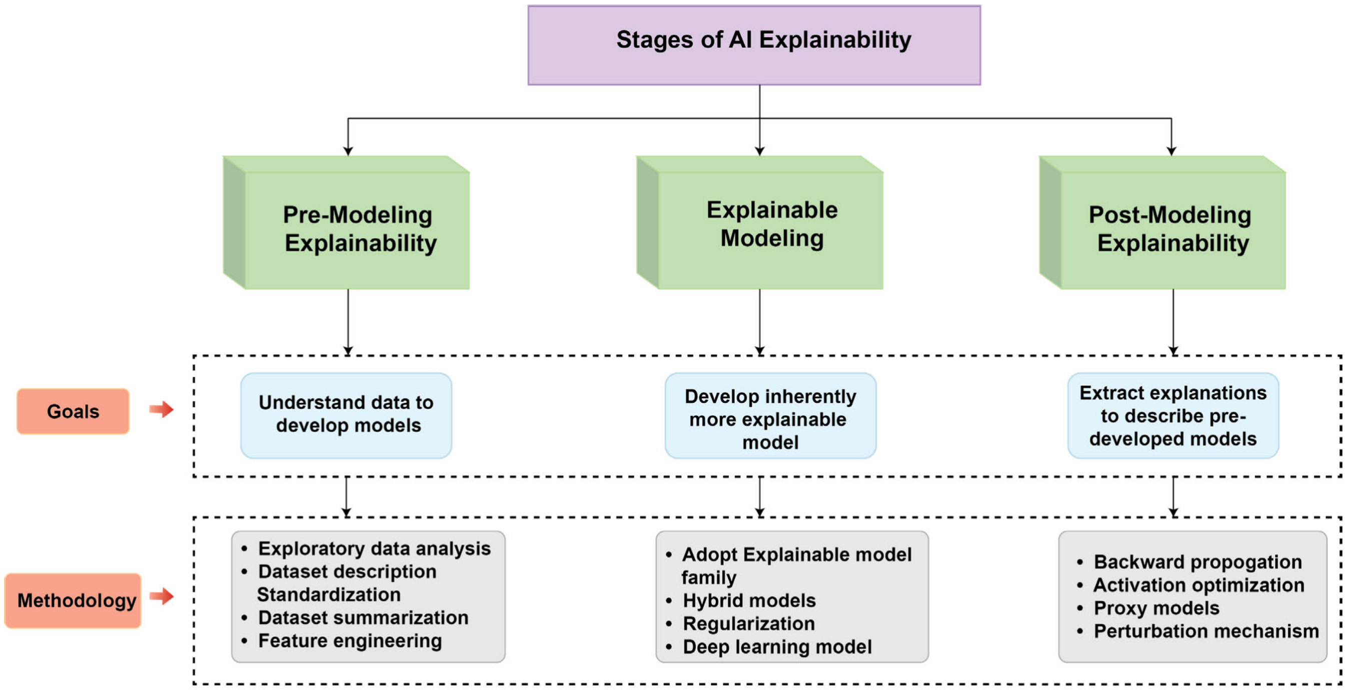

Stages of AI Explainability
Pre-Modeling Explainability
Explainable Modeling
Post-Modeling Explainability
Goals
Understand data to develop models
Develop inherently more explainable model
Extract explanations to describe pre-developed models
Methodology
• Exploratory data analysis
• Dataset description Standardization
• Dataset summarization
• Feature engineering
• Adopt Explainable model family
• Hybrid models
• Regularization
• Deep learning model
• Backward propogation
• Activation optimization
• Proxy models
• Perturbation mechanism

FIGURE 18. AI explainability stages.

for its working, for various input size. In simple words space it requires to complete the task. Space complexity in this case is n (n = no of websites). Gradient boosting machines are based on combining the predictions of numerous models to create a higher-quality model. To be more specific, models are iteratively trained using the mistake of previous models, giving priority to the more difficult cases. The errors are computed throughout each iteration, and a model is fitted to these mistakes. Finally, the contribution of each base model to the final model is determined by minimizing the overall error of the final model. To fit the basic models, the optimization technique called XGBoost is used.

XGBoost constructs the trees using the depth-first approach optimizes a gradient with parallel processing criteria and applies a regularization term penalty to prevent bias during training. A tree ensemble model illustrates the following for a given dataset of examples with m features (x, y).

$$Y = \sum_{k=1}^{k} f_k(x_i), f_k \in F \tag{16}$$

The space provided to regression trees is F. The following XGBoost uses a regularized goal to optimize and reduce the loss function.

$$I() = \sum_{i}^{k} Y_{true}, Y_{pred} + \sum_{i}^{k} \Omega(f_k) \tag{17}$$

$$\Omega(f_k) = \gamma * T + \frac{1}{2}\lambda|W|^2 \tag{18}$$

where γ*T is the complexity penalized term of the model and $\lambda|W|^2$ is the regularization term penalty.

C. EVALUATION

To evaluate and compare these prediction models, four criteria: Confusion matrix, Precision, recall and F1 score is used. These metrics can be calculated using confusion matrix parameters: true positive (the number of correctly categorized anomalous events); false positive (the number of typical events that are mistakenly categorized as abnormal); true negative (the number of normal instances that are correctly classified); and false-negative (the number of anomalous instances that are incorrectly classified as normal) and then receiver operating characteristic (ROC) plot and area under the curve (AUC) were plotted for each model.

a. *Confusion matrix* - The number of test records properly and wrongly predicted by a classification model is used to measure its performance. The confusion matrix indicates which a model successfully and wrongly predicts classes and in which errors are created. Recall, Precision, Specificity, Accuracy, and the AUC-ROC curve are all used to calculate it.

   True Positive (TP): When we predict an observation belongs to a class, and it does belong to that class. In this study, a website is classified as Credible and is Credible.

   True Negative (TN): When we predict an observation does not belong to a class and does not belong to it. In this study, a website is classified as not Credible and is not credible.

b. *Accuracy* – Number of correct predictions / Total number of predictions.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

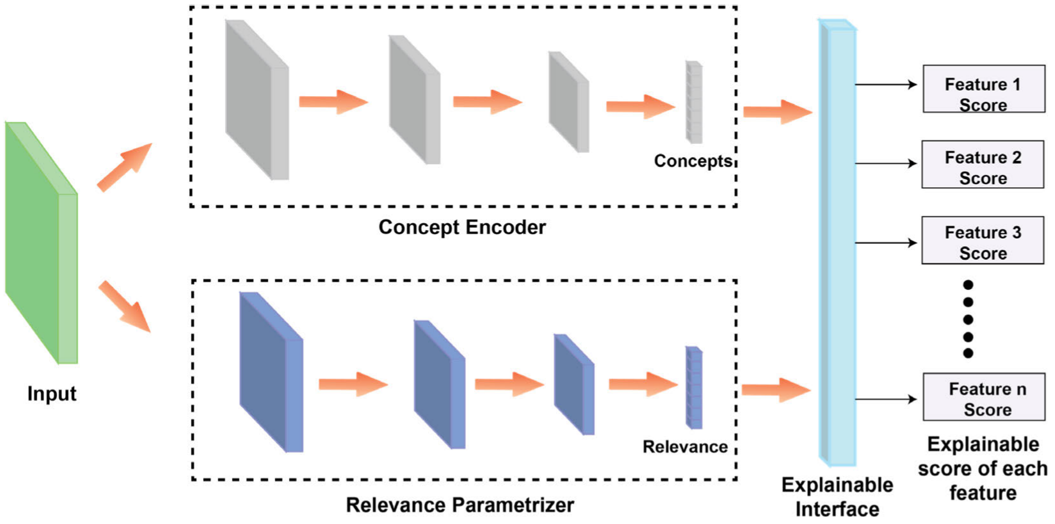


FIGURE 19. Shapley working.

c. *Precision* - Precision gives us the fraction of correctly identified as positive of all predicted as positives.

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

d. *Recall* – Recall gives us the fraction we correctly identified as positive out of all positives.

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

e. *F1 score* – It is defined as the harmonic mean of the model's precision and recall.

$$F1score = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \tag{22}$$

f. *ROC/AUC Curve* - A threshold can be used to convert probability outputs into classifications. By controlling a little bit of the confusion matrix, the receiver operator characteristic i plots the sensitivity and specificity for every possible decision-rule cutoff between 0 and 1. Thus, by altering the threshold, some of the numbers change in the confusion matrix. For every threshold, the ROC curve plots the False-positive rate and the True-positive rate.

D. SHAPLEY ADDITIVE EXPLANATIONS

Shapley Additive Explanations (SHAPs) [84], [85] are a local model-independent approach to analyzing predictions from a general Black-box model based on feature importance scores. The feature importance explanation describes how each type of input in the input influences the prediction. This type of analysis is common since there are many model-independent and model-specific ways of calculating the importance of local or global features.

Based on Shapley values, SHAP aims to explain the prediction function, for instance, xi, as a sum of its independent feature values. Here it is assumed that the individual feature values are in a cooperative game with an equal payout; Shapley values allow a fair distribution of the payout among the feature values.

In this paper, the Shapley values are obtained by the formula:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_{(x)}(S \cup \{i\}) - f_{(x)}] \tag{23}$$

In the above formula with f(x) features and M model, S represents each possible permutation of feature values except the i-th value. Here, |S|! indicates the number of feature values before the i-th value that is possible. (|M|-|S|-1) represents the number of features that follow the i-th feature value, the difference in the above equation being the marginal contribution to S of the added i-th feature. The above equation yields the values SHAP under the assumptions: $f(x_S) = E[f(x|x_S)]$. In other words, the prediction for any subset S of feature values is the expected value of the prediction for f(x) given the subset x_S.

The SHAPLEY formula aims to compute the contributions of each feature to the prediction of an instance x. So, to explain the predictions, the TreeShap model-type-specific approximation method is used. This method assumes feature independence and model linearity to simplify the computation of SHAP values. Using this method feature explanations were generated.

E. INTERNAL WORKING OF SHAPLEY

Shapley is a layered network where input, after passing through many layers, gives us desirable relevance and concepts. Input is given to the Shapley model, as shown in Fig (19). Then the concept-based encoders convert the input into a small number of interpretable features, and input-dependent parameters determine relevance scores.

Pseudocode 2 Tree Shap Pseudocode

```
Requirements: Classifier d, Input x of the explained prediction
Requirement: Distribution s of the training data are required.
Requirement: Least number of samples for each feature $F_{min}$
Requirement: Maximum number of samples for all the features $F_{max}$
for i = 1 to n do                                                    ▷ Initialization
    $m_i \leftarrow 0$
    .
    $\phi_i \leftarrow 0$
    .
while $\sum m_i < F_{max}$ do
    .
if $\exists i : m_i < F_{min}$ then
            pick feature j to be sampled s.t. $m_j < F_{min}$
    else
            choose the highest sampling rate for feature j. $\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!}[f_{(x)}(S \cup \{i\}) - f_{(x)}]$
return $\Phi_i$
```

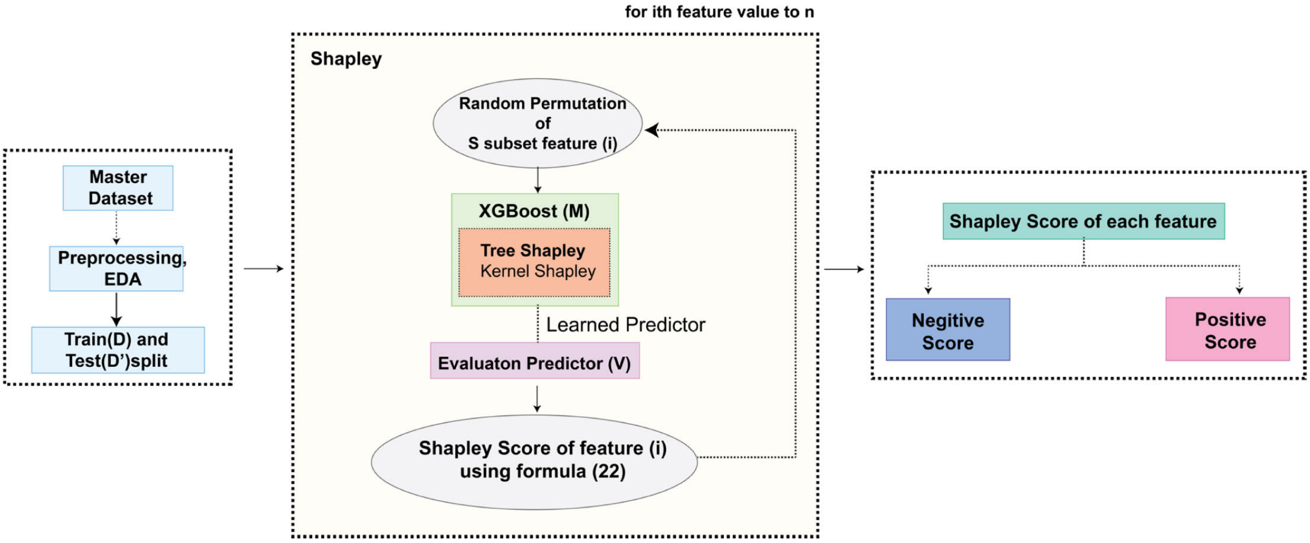

for ith feature value to n
Master Dataset
Preprocessing, EDA
Train(D) and Test(D')split
Shapley
Random Permutation of S subset feature (i)
XGBoost (M)
Tree Shapley
Kernel Shapley
Learned Predictor
Evaluaton Predictor (V)
Shapley Score of feature (i) using formula (22)
Shapley Score of each feature
Negitive Score
Positive Score

FIGURE 20. Shapley workflow.

To generate a prediction, an aggregation function is used. Locally, the full model behaves as a linear function with parameters producing both concepts and relevance interpretations that go to the final explainable interface and produce relevant explainable scores.

F. SHAPLEY VALUES

By considering pairwise attributions of features, SHAP allows us to compute interaction effects. The resulting matrix represents the impact of all feature pairs on a given prediction model. It depends on Shapley interaction index and is given by

$$\phi_{i,j} = \sum_{S \subseteq N \setminus (i,j)} \frac{|S|!(M - ||S| - 1)!}{M!} \Delta_{i,j}(S) \quad (24)$$

where,

$$\begin{aligned}\Delta_{i,j} &= f_x(S \cup \{i,j\}) - f_x(S \cup \{j\}) - [f_x(S \cup \{i\}) - f_x(S)] \\ &= f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S) \quad (25)\end{aligned}$$

Using the above equation, it can be seen that the SHAP interaction value of a particular feature in relation to another can be assessed as the difference between the SHAP values of a particular feature with & without a particular feature. Fig (20) represents the same.

As a result, the SHAP interaction between the i-th and j-th is split equally (i.e., $\Phi_{ij} = \Phi_{ji}$), and the total interaction is $\Phi_{ij} + \Phi_{ji}$. If a feature's SHAP value and the sum of SHAP interaction values are subtractions for the prediction, the main effect is obtained as follows:

$$\phi_{i,i} = \phi_i - \sum_{j \neq 1} \phi_{i,j} \quad (26)$$

VI. RESULTS

For any system, it is vital to examine its performance against several evaluation metrics to test how correctly it is generating the desired output. Also it is important to compare our model with the previous models. As seen in Fig (21), the Results section is divided into two modules:

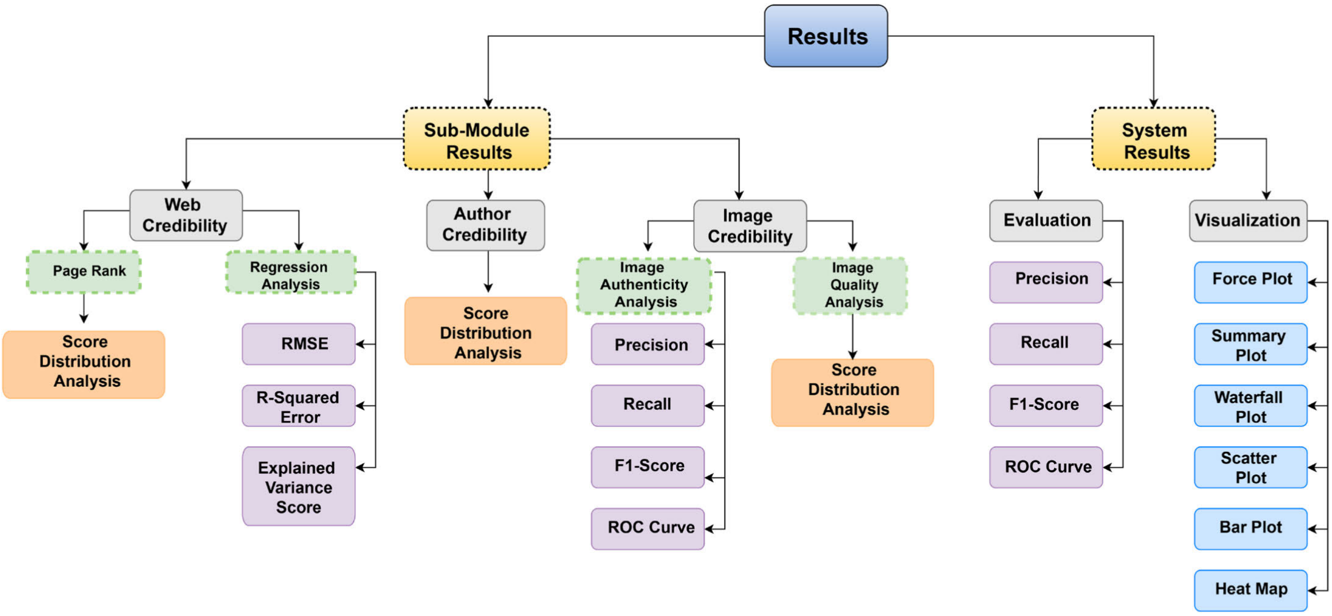


FIGURE 21. Results organization.

1) Sub-Module Results 2) System Results. The Sub Module Results section focuses on inspecting the outcomes of the modalities considered. We determine a suitable performance metric for each modality and its corresponding methodologies, subject our data and models to their respective tests, and finally, inspect the results. Finally, for the System Results section, we evaluate the performance of our system's results (XAI's performance) and provide suitable visualization aids to make it easier for readers to examine the work of our architecture. In the next few subsections, we discuss the results of each module and present the inferences we drew from each of them.

A. SUB MODULE-RESULTS

1) WEBSITE CREDIBILITY

a: PAGE RANK

After applying the Page Rank Algorithm, the page rank score of each web blog is derived. The obtained PageRank score between 0-100 is then scaled to 0-10 scale and stored them in the master dataset.

Websites with PageRank score 6 and above have more inbound links, sites with PageRank score 3 and Page Rank score 4 have fewer links, and news sites without any inbound links start at PageRank score 0. The distribution of these scores is shown below in Fig (22).

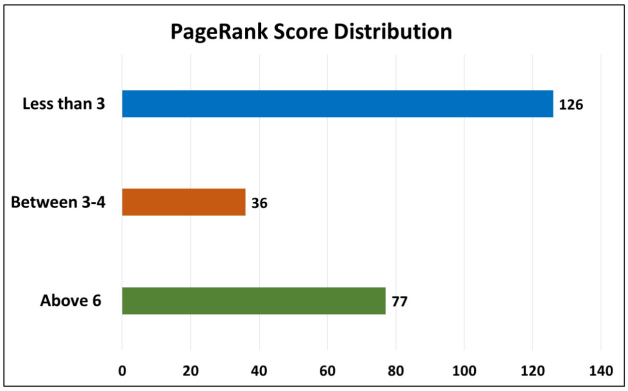


FIGURE 22. PageRank score distribution analysis.

b: REGRESSION ANALYSIS

In the regression analysis, machine-learning is used to correctly predict the platform credibility scores for each web blog in the dataset. A comparative analysis of the performance of these regression models with different evaluation criteria (R-squared Error, Root mean square error, and Explained variance score) is shown in Table 5. It is evident that the Decision tree regression performs admirably across all assessors, which can also be seen through the graphical representation of the evaluation results in Fig (23). The performance results of the regression models is shown in Table 5. It can be observed that the Decision tree has the highest R-squared error and explained variance score the lowest Root means square error. Consequently, these results can find the prediction scores of the website using decision tree regression and stored them in the master dataset as ***website_score.***

2) AUTHOR CREDIBILITY

After applying the four proposed tests to determine the author's writing style, a score for each of the tests and the overall credibility score of the author was obtained.

It was observed that most of the overall scores lay within the range of 5-8 on a scale of 0-10. When filtered as

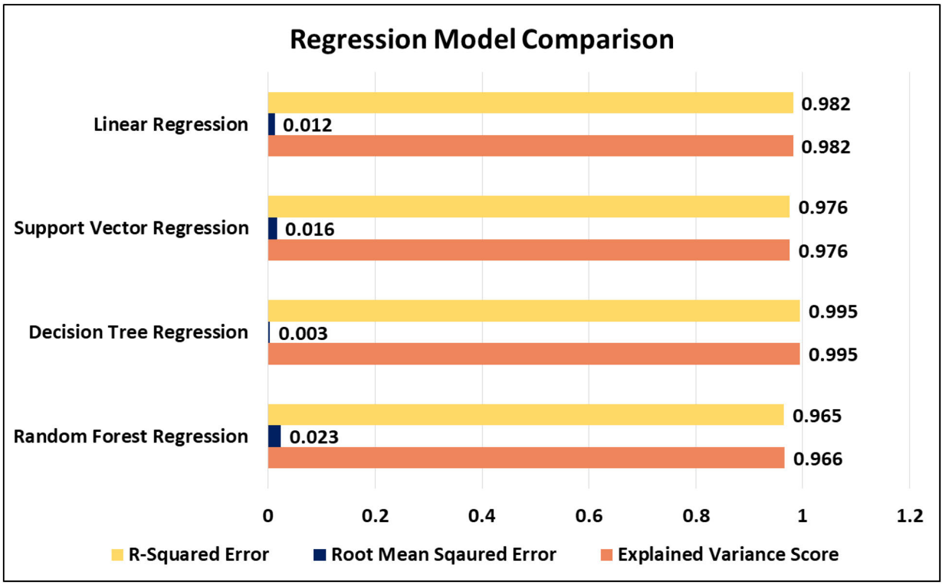


FIGURE 23. Regression models performance comparison.

TABLE 5. Regression model performance comparison.

Regression model	R-Squared score	Root mean Squared Error	Explained variance Score
Linear Regression	0.982	0.012	0.982
Support Vector Regression	0.976	0.016	0.976
Decision Tree Regression	**0.995**	**0.003**	**0.995**
Random Forest Regression	0.965	0.023	0.966

credible/noncredible blogs, it was observed that the non-credible blogs usually had a score of 4-6 while the credible ones had a score >6 with a peek between 8-8.5. Most of these health blogs tend to have a low domain expertise score, shifting its peak to a low 4-4.5 value only. This proves how most blogs on the internet are less likely to make use of proper medical terms when referencing treatments/remedies.

However, the Grammar & Typo Score remained in the higher ranges, proving that these blogs are often checked for correct spelling & grammar. The readability scores of most blogs lied in the range of 4-6, with a peak of 5.5, showing that most blogs are reader-friendly and easily understandable.

Other inferences drawn were i) Blogs with lower credibility have a significantly lower grammar score than their credible counterparts. ii) Readability score for both credible & noncredible blogs have less variance. iii) Credible blogs have a higher domain expertise score than non-credible blogs.

3) IMAGE CREDIBILITY

a: IMAGE AUTHENTICITY

To train the model, the authors used two approaches, one with data augmentation & another without it. The authors observed significant differences in the results of both these approaches. When subjected to data augmentation, the model had an average precision of 0.94, Recall of 0.93 & F1-Score of 0.94 (Table 6). This helps us conclude that the model does well when trained on a highly augmented dataset and correctly segregates credible images from their non-credible counterparts. However, when trained without augmentation, the dataset shows a precision, recall & f1-score of only 0.53 (Table 6). This proves that augmentation thus indeed improves the model's performance significantly, as shown in Fig (24). Augmentation helps add various flavors to the already existing dataset, subjecting the model to varying types of images to come across, making it more flexible when considering an image's authenticity.

b: IMAGE QUALITY

The authors successfully perform the image quality test on the images of the blogs. Each blog is assigned a quality image score in the range of 0-10. A high score indicates a better-quality image that can sustain various transformations & resizing, while an image with a lower score is not compatible with most platforms. It was also observed that most of the images have a low-quality score in the range of 2-4.

TABLE 6. Performance analysis (with and without data augmentation).

Metrics	With Augmentation Result	Without Augmentation Result
Avg Precision	0.94	0.53
Avg Recall	0.93	0.53
Avg F1-Score	0.94	0.53

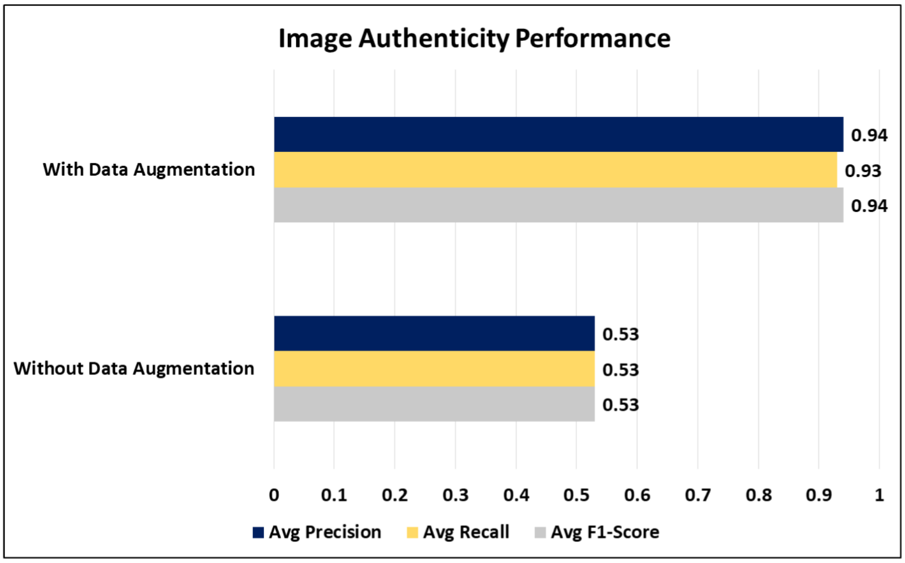


FIGURE 24. Image authenticity performance analysis.

TABLE 7. Performance comparison of classification models.

Classification Model	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0.938	0.934	1	0.966	0.973396
Decision Tree	0.925	0.945	0.972	0.958	0.763693
K-Nearest Neighbor	0.938	0.934	1	0.966	0.980438
XGBoost	**0.975**	**0.973**	**1**	**0.986**	**0.984351**

This quality score includes pixel distributions and feature distributions, indicating that most websites host images with low-quality indexes. This observation matches that most websites rarely prioritize the quality of the images they display on their blog since most of them are usually advertisements or remedy results etc. However, most credible blogs had images with quality scores as high as 8-9, which proves how these blogs focus on maintaining all aspects of their blogs, not just the literary or aesthetic appeals.

B. SYSTEM RESULTS

After subjecting the blogs in the master dataset to multiple classification models, in this section the authors analyze the performances of each of the models and select the most suitable model to be subjected to the SHAPLEY (explainable AI model) for feature explanations. The ROC–AUC curve of the proposed classification methods is shown in Fig (25). This figure shows that the XGboost model gives curves closer to the top-left corner and far from the 45 degrees diagonal of the Roc space, hence indicating better performance.

Table 7 And Fig (26) shows the comparison of different classification models using 5 evaluation parameters (Accuracy, Precision, Recall, F1-score, and AUC). From these results, the XGBoost model accomplishes the highest scores in every evaluation metric.

To summarize the results better, we compared the performance to previous existing techniques. Table 8 shows that the multimodal credibility analysis achieved highest accuracy in all aspects.

TABLE 8. Performance comparison of final model with previous models.

Modality	Modalities Used	Dataset	Accuracy	Precision	Recall	F1-score	AUC
Popat et al. (2019) : Text	Text	PolitiFact	0.6732	-	-	0.68	0.75
Blog Verified Content	Text	German APA News Corpus, Mined Text from Wikipedia (German,English, French)	-	0.83	-	-	-
Alexandra Olteanu	Web	Public Blog Features Dataset by Microsoft	0.68	0.68	0.68	0.68	-
Credibility-based Fake News Detection	Text	PolitiFact,Buzzfeed	-	-	-	0.80	
Information Credibility Twitter	Text	News on Twitter	-	0.754	0.742	0.739	-
Situala et al. (2019)[34]	Text and Web	PolitiFact, BuzzFeed	0.7-0.8	0.7-0.8	0.80	-	-
Helwe et al. (2019)[37]	Text and Web	Labeled Arabic Web blogs Dataset.	0.83	-	0.63	-	-
Olteanu et al. (2013)[50]	Text and Web	Public Blog Features Dataset by Microsoft	0.75	-	-	-	-
Proposed Multimodal Framework	**Text + Web +Image**	**Self-generated dataset**	**0.975**	**0.973**	**1**	**0.986**	**0.984351**

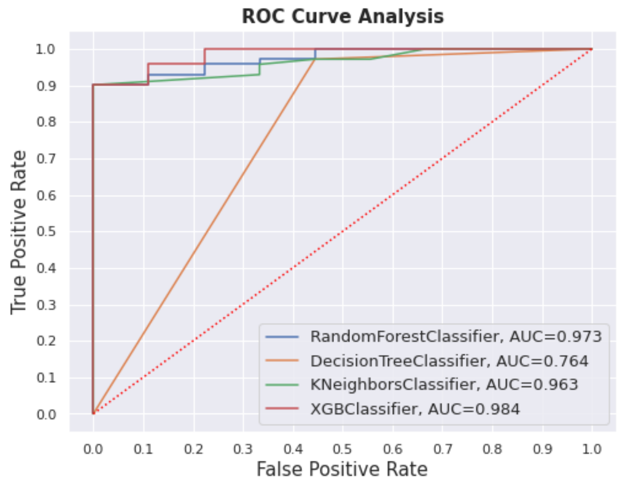


FIGURE 25. ROC curve for performance analysis of classification models.

1) EXPLAINING MODEL DECISIONS

This section uses SHAPLEY to explain why Blogs are classified as credible or not credible by presenting feature importance through various Shapley plots. Multiple Shapley plots are visualized to have a better understanding of the features discussed below.

a: FORCE PLOT

A prediction's output value is generated from the sum of the base value (average prediction over the validation set). A feature attribute such as Shapley value can be viewed as a "force"; a feature's cost affects the prediction in either direction. The prediction is obtained from the baseline. Each Shapley value represents a force that pushes either to increase (positive value) or decrease (negative value) a prediction in the plot. These effects balance each other at the actual prediction of data instance. Fig (27) shows SHAP explanation force plots for two Websites from the master dataset collected using each credibility score.

The base value is 2.28, and the model projected 2.55. Features that increase the prediction accuracy are depicted in pink. The amount of the feature's effect is shown by their visual size. Feature values that reduce the prediction's accuracy are shown in blue. The Page rank score is 2 and has the most positive impact. The Image score value, on the other hand, has a significant decreasing impact on the forecast.

Fig (28) shows the feature importance of multiple websites (say 50) taken at a point in time. Fig (28 a) and (28 b) show the possible outcomes that can be chosen to visualize the plot through different aspects. Here y-axis shows the contribution value, and the x-axis shows the number of websites.

Force plot of multiple websites

b: SUMMARY PLOT

The Summary plot summarizes feature importance with its associated feature effects. An indicator plot displays the Shapley values for each feature and instance. On the y-axis, a feature determines position, and the value of Shapley determines the position on the x-axis. In the color wheel, low to

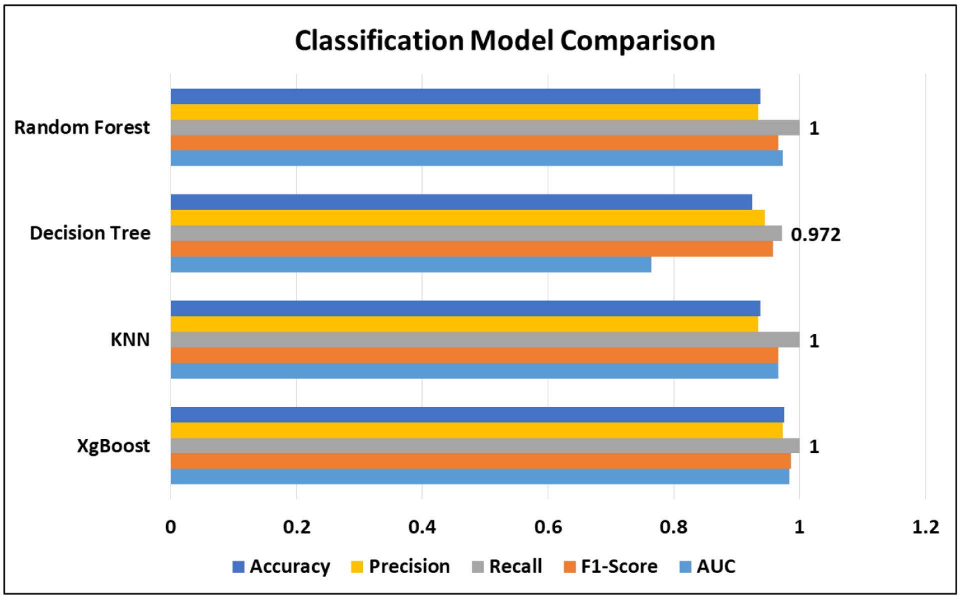


FIGURE 26. Performance comparison of classification model.

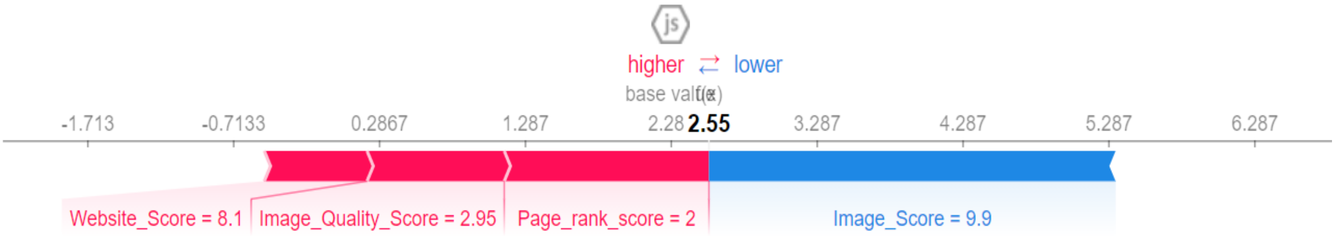


FIGURE 27. Force plot visualization.

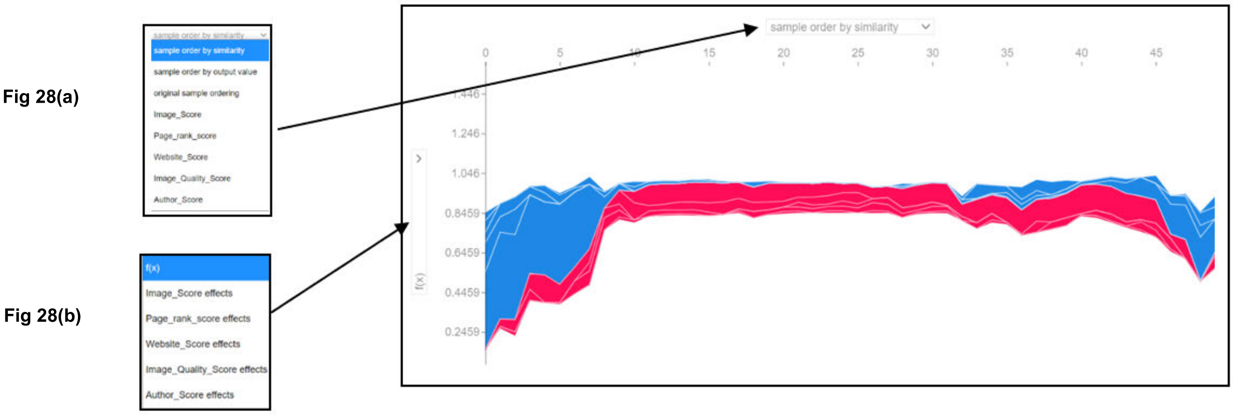


FIGURE 28. Force plot for multiple websites.

high values correspond to the features. With the overlapping points jittered in the y-axis direction, it is possible to get a sense of the distribution of the Shapley values. They are ranked from most important to least important in Fig (29).

This plot illustrates the following information:

- *Feature importance*: Variables are sorted by decreasing importance.
- *Impact*: By looking at the horizontal location, it is possible to see whether the effect of that value leads to a lower or higher prediction.
- *Original value*: A color identifies how high or low that variable is for that observation depending on its value.
- *Correlation*: The "Image_score" has a high and negative impact on the overall score. The "high" comes from

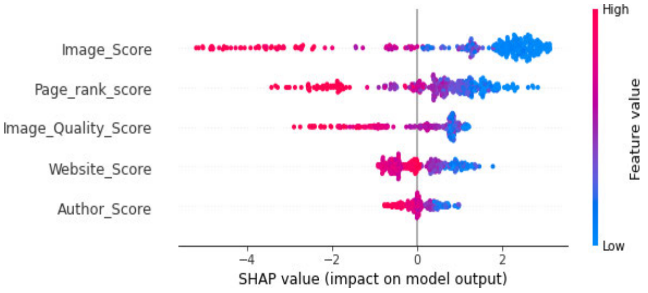


FIGURE 29. Summary plot visualization.

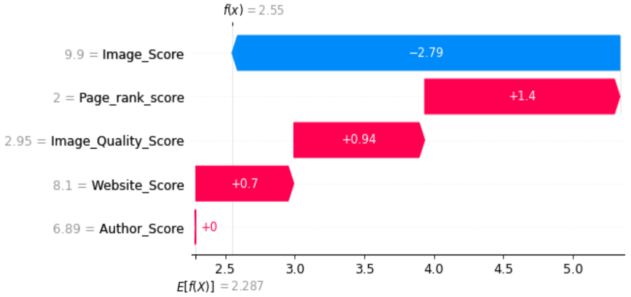


FIGURE 30. Waterfall plot visualization.

the red color, and the "negative" impact is shown on the X-axis.

c: WATERFALL PLOT

The waterfall plot strongly explains why a case receives the prediction; it does give its variable values. It demonstrates how each feature contributes to pushing the model output from the base value to its prediction by adding (red) or deleting (blue) the values to achieve the final prediction. Below is the graph for the first observation in X_test. The average of all observations is calculated as the base value of 2.287 at the bottom in Fig (30). The final prediction for this observation (on top) is 2.55+1.4+0.94+0.7 (note the rounding error). Each variable's value is next to its name, e.g., the value for "Image_Score" on the first observation is 9.9.

d: SCATTER PLOT

To acquire a clear idea of the impact a feature has on the output of a model, the SHAP value of a feature versus the value of the feature for all the cases in a dataset can be plotted.

Since SHAP values represent a feature's contribution to the change in the output, the figure below displays the change in Website Score (Regression Score) in Fig (31). Short vertical lines represent the impacts of feature interaction.

and the scatter plot can help to determine the best feature to color if the entire explanation tensor is passed. Image Score is selected in this case.

e: BAR PLOT

By passing a matrix of SHAP values to the bar plot function, a plot of global feature importance is created with the mean absolute value for each feature over all the given samples. So to get an explanation with many samples, the authors plotted the mean absolute value for each feature column as a bar chart Fig (32).

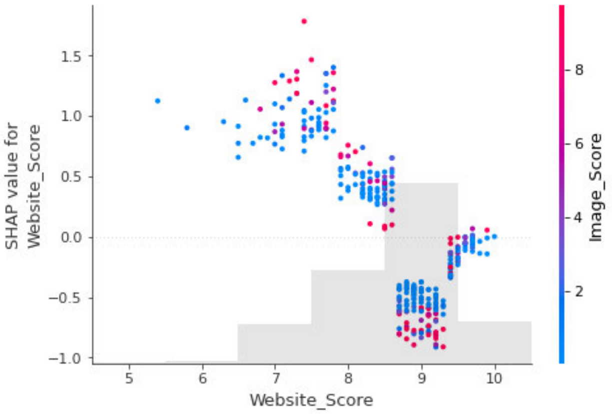


FIGURE 31. Scatter plot visualization.

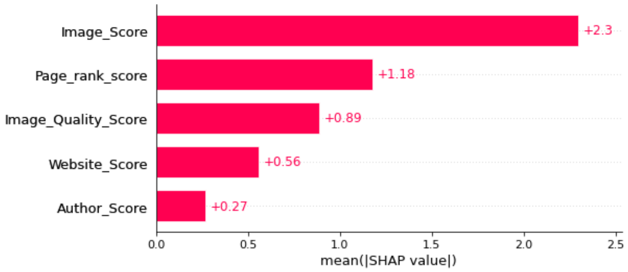
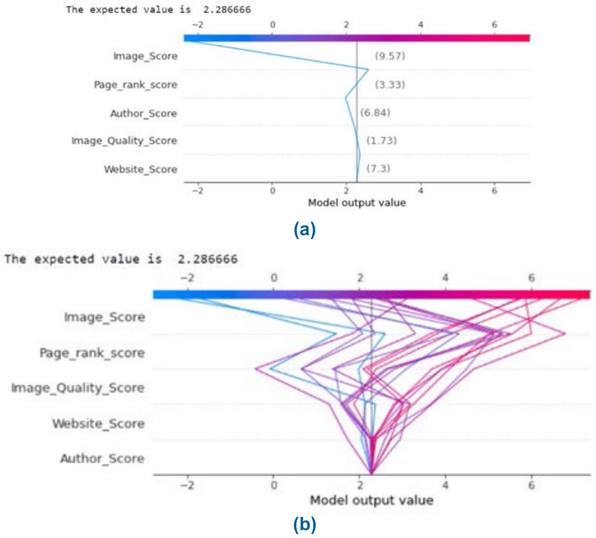


FIGURE 32. Bar plot visualization.

It does not talk more about the increase or decrease of the feature for prediction; rather just explains how features affect credibility prediction. Therefore, the study can specify how many predictors to display, and can also sum up the tail predictors. By doing so, it is possible to inform the audience about the collective contribution of the tail predictors. A typical bar plot may be created with merely the mean absolute value of SHAP values for each feature shown in Fig (32).

f: DECISION PLOT

The vertical line in the decision plot marks the baseline of the model. Colored lines depict predictions. The prediction line is printed along with the feature value for reference. At the bottom of the plot, the prediction line represents how the SHAP values (i.e., the feature effects) accumulate from the base score to achieve the model's final score at the top. (This works similarly to a linear statistical model, in which the sum of effects plus the intercept is the prediction.) Decision plots provide a simple definition of SHAP values. Fig (33a) shows the visualization of 1st website, while Fig (33b) shows the visualization of multiple websites (20). The model's output is shown on the x-axis. The odds here are in logarithmic form. Explainer expected_value is plotted along the x-axis. As with linear models, SHAP values are



(a)

(b)

FIGURE 33. (a) Decision plot of a single website. (b) Decision plot of multiple websites.

compared to the linear model's expected value. Features are listed along the y-axis. Features are by default sorted in ascending importance. A plot of observations is used to calculate the importance. Feature importance ordering, hierarchical cluster feature ordering, and user-defined feature ordering are all supported by the decision plot. The decision plot can enable cluster feature ordering and user-defined feature ordering in addition to generic feature importance ordering. Colored lines represent observational forecasts. Each line on the plot crosses the x-axis at the value predicted for its corresponding observation. The value of this parameter defines the color of a line on a spectrum. In addition to the model's base value, SHAP values are added to each feature from the bottom to the top. The overall prediction is based on how each feature contributes. Observations convergence at explainer.expected_value, on the bottom of the plot.

g: HEAT MAP PLOT

To better explore interactions, a heatmap can be very useful. The heatmap plot function enables us to generate a plot with the instances on the x-axis, the model inputs on the y-axis, and the SHAP values encoded as colors. It demonstrates how each pixel affects the prediction. Using a heatmap, it can be observed how much data is present in two dimensions Fig (34). Depending on how the colors differ, it determined how the data are clustered. The hot-to-cold color scheme is used to illustrate the relationship. Y-axis bars corresponding to the bars in a bar chart are displayed in descending order, with bars representing variable importance.

On top is the f(x) curve, showing how the model predicts the instances. Upon clustering instances, the SHAP orders them (using shap.order.hclust) on the X-axis. Based on the 2D heatmap data, the base_value (using.base_value) is the mean prediction overall times. It can be seen that the importance of each feature in every website through the pixels. For, e.g., Image_score has a dark blue pixel from the 25th-35th website, indicating that Image_Score decreases the prediction in these websites. It is a very useful Shapley plot as it gives the inference of features in all the websites, which makes users find the contribution of each website's features in one plot.

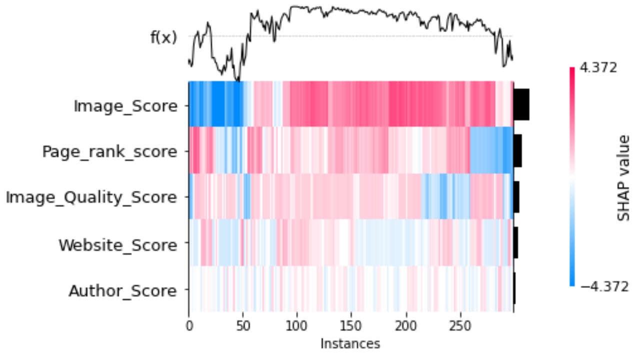


FIGURE 34. Heat map to explore interactions.

VII. DISCUSSION

This Multimodal approach to find the credibility of health blogs and then explaining the predictions in a user-understandable format provides valuable and vital information. From an exhaustive literature review, most of the work in the field of credibility is done only for news articles and political facts using only the website & text of the blog as the modes of analysis. Even though the initial step of finding the methods that make users understand weblogs' credibility in every aspect was time-consuming, the overall results were satisfying and significant. The main contribution of this study is the credibility assessment of the blog is made not only on the basis of a single modality which can lead to some kind of bias. Multiple modalities can help more accurate creditability assessment. Also, this work provides software-based explanations about the predictions made. As per authors' knowledge very less similar such work has been carried out in healthcare domain. However, the authors also faced certain limitations & challenges as follows.

1. Limited work on multi-modal credibility techniques posed a serious challenge during the initial research work of this project.
2. All the datasets for this project had to be created using various web-scraping tools because of the lack of publicly available datasets.
3. This project's scope is limited to only analyzing the authors' writing style but not verifying their claims.
4. The techniques in this project have been applied to only health blogs hosted online.

VIII. FUTURE WORK

While the work examines credibility from three major aspects & provides results with suitable explanations, there is scope for further improvement & advancements as well. The study has focused strictly on the health domain while scrutinizing credibility, the same framework can be applied to different domains like entertainment or political news, and even technological advancements & information. Domain adaptation [86]–[88] can be explored, where the health domain could serve as the ‘‘source’’ domain and be applied to different ‘‘target’’ domains listed above [89]. Our multi-modal credibility analysis framework could thus achieve a universal approach. Our focus for this project has purely been blogs hosted online. This could be extended to other communications mediums, extending but not limiting social media content sharing (Twitter, Instagram) [90], [91]. Another scope could be extended to examining the sharing networks of authors. Usually, credible/ correct news authors collaborate amongst themselves, while the same is observed in fake news authors. A systematic review of these co-authorship & content sharing networks could help extend the results of this project. Lastly, this project only focused on analyzing credibility by examining the author's writing style & not by verifying the actual authenticity of their claims [92], [93]. This could serve as an additional feature of this framework in the future.

IX. CONCLUSION

In this study the authors have proposed an Explainable AI assisted multimodal credibility analysis framework that helps to classify health misinformation in online beauty blogs. The framework does the credibility assessment via Web, Image and Author information. The study generates an overall credibility score for each blog. Setting a threshold of 6, the study classifies each blog as credible or non-credible. For this purpose, an amalgamation of various analysis tools and AI models have been utilized and an accuracy of 97.5% is achieved. Then using the SHAPLEY tool, the framework provides suitable explanations of the final classification model's decisions, with the help of multiple visualization aids. This system can be very useful for people, especially teens, to better understand a credibility of a beauty blog suggestions and take an informed decision about following those tips. Our system, upon integration with other domains or fields such as healthcare, finance, pandemic management can also produce useful and significant results of representing information credibility.

DATA AVAILABILITY

The code and data for this research is available at GitHub: https://github.com/vidsssw/Explainable-AI-for-Multimodal-Credibility-Analysis-of-Online-Beauty-Health-Mis–Information.

REFERENCES

[1] J. Morahan-Martin and C. D. Anderson, ‘‘Information and misinformation online: Recommendations for facilitating accurate mental health information retrieval and evaluation,’’ *CyberPsychol. Behav.*, vol. 3, no. 5, pp. 731–746, Oct. 2000.

[2] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand. (2019). *Understanding and Reducing the Spread of Misinformation Online*. [Online]. Available: https//psyarxiv.com/3n9u8

[3] L. D. Scherer and G. Pennycook, ‘‘Who is susceptible to online health misinformation?’’ Amer. Public Health Assoc., Washington, DC, USA, Tech. Rep. S276-S277, 2020.

[4] X. Zhang and A. A. Ghorbani, ‘‘An overview of online fake news: Characterization, detection, and discussion,’’ *Inf. Process. Manage.*, vol. 57, no. 2, Mar. 2020, Art. no. 102025.

[5] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, ‘‘The spreading of misinformation online,’’ *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 3, pp. 554–559, 2016.

[6] J. Alexander and J. Smith, ‘‘Disinformation: A taxonomy,’’ *IEEE Secur. Privacy Mag.*, vol. 9, no. 1, pp. 58–63, Jan. 2011, doi: 10.1109/MSP.2010.141.

[7] D. Paschalides, A. Kornilakis, C. Christodoulou, R. Andreou, G. Pallis, M. Dikaiakos, and E. Markatos, ‘‘Check-it: A plugin for detecting and reducing the spread of fake news and misinformation on the web,’’ in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Oct. 2019, pp. 298–302.

[8] M.-A. Abbasi and H. Liu, ‘‘Measuring user credibility in social media,’’ in *Proc. Int. Conf. Social Comput., Behav.-Cultural Modeling, Predict.*, 2013, pp. 441–448.

[9] R. Fletcher, A. Cornia, L. Graves, and R. K. Nielsen, ‘‘Measuring the reach of 'fake news' and online disinformation in Europe,’’ *Australas. Policing*, vol. 10, no. 2, 2018.

[10] W. H. Li, ‘‘Detecting non-credible news using machine learning,’’ Tech. Rep., 2018.

[11] S. Akamine, Y. Kato, K. Inui, and S. Kurohashi, ‘‘Using appearance information for web information credibility analysis,’’ in *Proc. 2nd Int. Symp. Universal Commun.*, Dec. 2008, pp. 363–365, doi: 10.1109/ISUC.2008.80.

[12] A. A. Shah, S. D. Ravana, S. Hamid, and M. A. Ismail, ‘‘Web pages credibility scores for improving accuracy of answers in web-based question answering systems,’’ *IEEE Access*, vol. 8, pp. 141456–141471, 2020, doi: 10.1109/ACCESS.2020.3013411.

[13] D. Kim and T. J. Johnson, ‘‘A shift in media credibility: Comparing internet and traditional news sources in South Korea,’’ *Int. Commun. Gazette*, vol. 71, no. 4, pp. 283–302, Jun. 2009.

[14] K. Shu, S. Wang, D. Lee, and H. Liu, ‘‘Mining disinformation and fake news: Concepts, methods, and recent advancements,’’ in *Disinformation, Misinformation, and Fake News in Social Media*. Cham, Switzerland: Springer, 2020, pp. 1–19.

[15] Y. Parfenenko, A. Verbytska, D. Bychko, and V. Shendryk, ‘‘Application for medical misinformation detection in online forums,’’ in *Proc. Int. Conf. e-Health Bioeng. (EHB)*, Oct. 2020, pp. 1–4, doi: 10.1109/EHB50910.2020.9280120.

[16] Y. Liu, K. Yu, X. Wu, L. Qing, and Y. Peng, ‘‘Analysis and detection of health-related misinformation on Chinese social media,’’ *IEEE Access*, vol. 7, pp. 154480–154489, 2019, doi: 10.1109/ACCESS.2019.2946624.

[17] M. Baildon and J. S. Damico, ‘‘How do we know: Students examine issues of credibility with a complicated multimodal web-based text,’’ *Curriculum Inquiry*, vol. 39, no. 2, pp. 265–285, Mar. 2009.

[18] V. K. Singh, I. Ghosh, and D. Sonagara, ‘‘Detecting fake news stories via multimodal analysis,’’ *J. Assoc. Inf. Sci. Technol.*, vol. 72, no. 1, pp. 3–17, Jan. 2021.

[19] N. Saini, M. Singhal, M. Tanwar, and P. Meel, ‘‘Multimodal, semi-supervised and unsupervised web content credibility analysis frameworks,’’ in *Proc. 4th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2020, pp. 948–955.

[20] G. Garzone, ‘‘Multimodal analysis,’’ *Handbook Bus. Discourse*, pp. 155–165, 2009.

[21] W. Samek, T. Wiegand, and K.-R. Müller, ‘‘Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,’’ 2017, *arXiv:1708.08296*. [Online]. Available: http://arxiv.org/abs/1708.08296

[22] S. Tseng and B. Fogg, ‘‘Credibility and computing technology,’’ *Commun. ACM*, vol. 42, no. 5, pp. 39–44, 1999.

[23] S. Y. Rieh and D. R. Danielson, ‘‘Credibility: A multidisciplinary framework,’’ *Annu. Rev. Inf. Sci. Technol.*, vol. 41, no. 1, pp. 307–364, 2007, doi: 10.1002/aris.2007.1440410114.

[24] S. M. Shariff, ‘‘A review on credibility perception of online information,’’ in *Proc. 14th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2020, pp. 1–7, doi: 10.1109/IMCOM48794.2020.9001724.

[25] N. C. Burbules, ‘‘Paradoxes of the web: The ethical dimensions of credibility,’’ Tech. Rep., 2001.

[26] V. L. Rubin and E. Liddy, "Assessing credibility of weblogs," in *Proc. AAAI Spring Symp., Comput. Approaches Analyzing Weblogs*, 2006, pp. 187–190.

[27] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Tech. Rep., 1999.

[28] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *J. ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1999, doi: 10.1145/324133.324140.

[29] X. Zhou and R. Zafarani, "A survey of fake news," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–40, Oct. 2020, doi: 10.1145/3395046.

[30] J. Abernethy, O. Chapelle, and C. Castillo, "Graph regularization methods for web spam detection," *Mach. Learn.*, vol. 81, no. 2, pp. 207–225, Nov. 2010, doi: 10.1007/s10994-010-5171-1.

[31] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," in *Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1998, pp. 104–111, doi: 10.1145/290941.290972.

[32] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. Int. Conf. Web Search Web Data Mining (WSDM)*, 2008, pp. 219–230, doi: 10.1145/1341531.1341560.

[33] N. Sitaula, C. K. Mohan, J. Grygiel, X. Zhou, and R. Zafarani, "Credibility-based fake news detection," Tech. Rep., 2020.

[34] W. Jaworski, E. Rejmund, and A. Wierzbicki, "Credibility microscope: Relating web page credibility evaluations to their textual content," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI), Intell. Agent Technol. (IAT)*, Aug. 2014, pp. 297–302, doi: 10.1109/WI-IAT.2014.47.

[35] R. M. B. Al-Eidan, H. S. Al-Khalifa, and A. S. Al-Salman, "Towards the measurement of Arabic weblogs credibility automatically," in *Proc. 11th Int. Conf. Inf. Integr. Web-Based Appl. Services (iiWAS)*, 2009, pp. 618–622, doi: 10.1145/1806338.1806455.

[36] C. Helwe, S. Elbassuoni, A. Al Zaatari, and W. El-Hajj, "Assessing Arabic weblog credibility via deep co-learning," in *Proc. 4th Arabic Natural Lang. Process. Workshop*, 2019, pp. 130–136, doi: 10.18653/v1/W19-4614.

[37] R. Manjula and M. S. Vijaya, "Deep neural network for evaluating web content credibility using Keras sequential model," in *Advances in Electrical and Computer Technologies*, 2020.

[38] K. Popat, "Credibility analysis of textual claimswith explainable evidence," Ph.D. dissertation, Universität des Saarlandes, Saarbrücken, Germany, 2019.

[39] A. Juffinger, M. Granitzer, and E. Lex, "Blog credibility ranking by exploiting verified content," in *Proc. 3rd Workshop Inf. Credibility Web (WICOW)*, 2009, pp. 51–58, doi: 10.1145/1526993.1527005.

[40] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep fake image detection based on pairwise learning," *Appl. Sci.*, vol. 10, no. 1, p. 370, Jan. 2020, doi: 10.3390/app10010370.

[41] J. Kocić, I. Popadić, and B. Livada, "Image quality parameters: A short review and applicability analysis," in *Proc. 7th Int. Sci. Conf. Defensive Technol.*, May 2016, pp. 1–6.

[42] B. Al-Duwairi, I. Khater, and O. Al-Jarrah, "Detecting image spam using image texture features," *Int. J. Inf. Secur. Res.*, vol. 3, no. 4, pp. 344–353, Dec. 2013, doi: 10.20533/ijisr.2042.4639.2013.0040.

[43] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[44] Z. Jin, J. Cao, J. Luo, and Y. Zhang, "Image credibility analysis with effective domain transferred deep networks," 2016, *arXiv:1611.05328*. [Online]. Available: http://arxiv.org/abs/1611.05328

[45] Y. Yamamoto and K. Tanaka, "ImageAlert: Credibility analysis of text-image pairs on the web," in *Proc. ACM Symp. Appl. Comput. (SAC)*, 2011, pp. 1724–1731, doi: 10.1145/1982185.1982546.

[46] S. Middleton, "Extracting attributed verification and debunking reports from social media: MediaEval-2015 trust and credibility analysis of image and video," Tech. Rep., May 2015.

[47] C. Shen, M. Kasra, W. Pan, G. A. Bassett, Y. Malloch, and J. F. O'Brien, "Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online," *New Media Soc.*, vol. 21, no. 2, pp. 438–463, Feb. 2019, doi: 10.1177/1461444818799526.

[48] N. Choudhary, R. Singh, I. Bindlish, and M. Shrivastava, "Neural network architecture for credibility assessment of textual claims," May 2018, *arXiv:1803.10547*. [Online]. Available: http://arxiv.org/abs/1803.10547

[49] A. Olteanu, S. Peshterliev, X. Liu, and K. Aberer, "Web credibility: Features exploration and credibility prediction," in *Proc. Eur. Conf. Inf. Retr.*, 2013, pp. 557–568.

[50] F. J. C. Garcia, D. A. Robb, X. Liu, A. Laskov, P. Patron, and H. Hastie, "Explain yourself: A natural language interface for scrutable autonomous robots," 2018, *arXiv:1803.02088*. [Online]. Available: http://arxiv.org/abs/1803.02088

[51] P. Hall, N. Gill, and N. Schmidt, "Proposed guidelines for the responsible use of explainable machine learning," 2019, *arXiv:1906.03533*. [Online]. Available: http://arxiv.org/abs/1906.03533

[52] D. Holliday, S. Wilson, and S. Stumpf, "User trust in intelligent systems: A journey over time," in *Proc. 21st Int. Conf. Intell. User Interfaces*, Mar. 2016, pp. 164–168.

[53] M. Harbers, K. van den Bosch, and J.-J. Meyer, "Design and evaluation of explainable BDI agents," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, Aug. 2010, pp. 125–132.

[54] F. Yang, S. K. Pentyala, S. Mohseni, M. Du, H. Yuan, R. Linder, E. D. Ragan, S. Ji, and X. Hu, "XFake: Explainable fake news detector with visualizations," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 3600–3604.

[55] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable fake news detection," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 395–405, doi: 10.1145/3292500.3330935.

[56] S. Mohseni, F. Yang, S. Pentyala, M. Du, Y. Liu, N. Lupfer, X. Hu, and S. Ji, "Trust evolution over time in explainable AI for fake news detection," Tech. Rep., 2020.

[57] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Explainable machine learning for fake news detection," in *Proc. 10th ACM Conf. Web Sci. (WebSci)*, 2019, pp. 17–26, doi: 10.1145/3292522.3326027.

[58] M. Chen, N. Wang, and K. P. Subbalakshmi, "Explainable rumor detection using inter and intra-feature attention networks," May 2020, *arXiv:2007.11057*. [Online]. Available: http://arxiv.org/abs/2007.11057

[59] Y.-J. Lu and C.-T. Li, "GCAN: Graph-aware co-attention networks for explainable fake news detection on social media," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 505–514, doi: 10.18653/v1/2020.acl-main.48.

[60] P. Ledin and D. Machin, *Introduction to Multimodal Analysis*. London, U.K.: Bloomsbury Publishing, 2020.

[61] J. Bezemer and C. Jewitt, "Multimodal analysis: Key issues," *Res. Methods Linguist.*, vol. 180, Apr. 2010.

[62] S. García, A. Fernández, and F. Herrera, "Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems," *Appl. Soft Comput.*, vol. 9, no. 4, pp. 1304–1314, 2009.

[63] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[64] R. Flesch, "Flesch–Kincaid readability test," Tech. Rep., Oct. 2007, vol. 26, no. 3.

[65] M. J. Peterson, "Comparison of Flesch readability scores with a test of reading comprehension," *J. Appl. Psychol.*, vol. 40, no. 1, p. 35, 1956.

[66] E. Charniak, "Statistical parsing with a context-free grammar and word statistics," in *Proc. AAAI/IAAI*, 1997, vol. 2005, nos. 598–603, p. 18.

[67] L. Kovacs and P. Barabas, "Experiences in building of context-free grammar tree," in *Proc. IEEE 9th Int. Symp. Appl. Mach. Intell. Informat. (SAMI)*, Jan. 2011, pp. 67–71, doi: 10.1109/SAMI.2011.5738850.

[68] J. Chaki and N. Dey, *A Beginner's Guide to Image Preprocessing Techniques*. Boca Raton, FL, USA: CRC Press, 2018.

[69] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[70] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*. [Online]. Available: http://arxiv.org/abs/1712.04621

[71] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. PhD Workshop (IIPhDW)*, May 2018, pp. 117–122.

[72] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Netw. Vis. Recognit.*, vol. 11, pp. 1–8, 2017.

[73] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[74] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, pp. 1–40, Dec. 2016.

[75] H. Ravishankar, P. Sudhakar, R. Venkataramani, S. Thiruvenkadam, P. Annangi, N. Babu, and V. Vaidya, "Understanding the mechanisms of deep transfer learning for medical images," in *Deep Learning and Data Labeling for Medical Applications*. Cham, Switzerland: Springer, 2016, pp. 188–196.

[76] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[77] M. Sandilya and S. R. Nirmala, "Determination of reconstruction parameters in compressed sensing MRI using BRISQUE score," in *Proc. Int. Conf. Inf., Commun., Eng. Technol. (ICICET)*, Aug. 2018, pp. 1–5.

[78] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[79] C. Yang, X. Zhang, P. An, L. Shen, and C.-C.-J. Kuo, "Blind image quality assessment based on multi-scale KLT," *IEEE Trans. Multimedia*, vol. 23, pp. 1557–1566, 2021, doi: 10.1109/TMM.2020.3001537.

[80] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," 2016, *arXiv:1605.01713*. [Online]. Available: http://arxiv.org/abs/1605.01713

[81] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[82] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy," 2017, *arXiv:1701.08317*. [Online]. Available: http://arxiv.org/abs/1701.08317

[83] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, A. C. Thanopoulou, and K. S. Nikita, "An explainable XGBoost-based approach towards assessing the risk of cardiovascular disease in patients with type 2 diabetes mellitus," in *Proc. IEEE 20th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2020, pp. 859–864, doi: 10.1109/BIBE50027.2020.00146.

[84] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using improved Shapley additive explanation," in *Proc. 10th ACM Int. Conf. Bioinf., Comput. Biol. Health Informat.*, Sep. 2019, p. 546.

[85] S. Park, J. Moon, and E. Hwang, "Explainable anomaly detection for district heating based on Shapley additive explanations," in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2020, pp. 762–765, doi: 10.1109/ICDMW51313.2020.00111.

[86] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[87] W. Mei and D. Weihong, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Jul. 2018.

[88] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2007, p. 137.

[89] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, "BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[90] M. AlRubaian, M. Al-Qurishi, M. Al-Rakhami, M. M. Hassan, and A. Alamri, "CredFinder: A real-time tweets credibility assessing system," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 1406–1409, doi: 10.1109/ASONAM.2016.7752431.

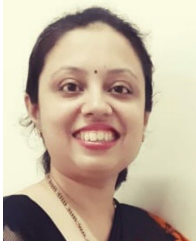
[91] M. Wijesekara and G. U. Ganegoda, "Source credibility analysis on Twitter users," in *Proc. Int. Res. Conf. Smart Comput. Syst. Eng. (SCSE)*, Sep. 2020, pp. 96–102, doi: 10.1109/SCSE49731.2020.9313064.

[92] N. Yarrabelly and K. Karlapalem, "Estimating credibility of news authors from their WIKI validated predictions," *NewsIR@ ECIR*, vol. 2079, pp. 12–17, 2018.

[93] M. M. U. Rony, E. Hoque, and N. Hassan, "ClaimViz: Visual analytics for identifying and verifying factual claims," in *Proc. IEEE Visualizat. Conf. (VIS)*, Oct. 2020, pp. 246–250.

VIDISHA WAGLE is currently working at Microsoft India Pvt., Ltd. She was also a Research Intern at the Symbiosis Centre for Applied Artificial Intelligence (SCAAI). Her research interests include artificial intelligence and machine learning domain, which includes areas like natural language processing, misinformation detection, multimodal deep learning, credibility analysis studies, and explainable AI.

KULVEEN KAUR is currently working at Amazon Development Centre India Pvt., Ltd. Her previous field placement was with the Symbiosis Centre for Applied Artificial Intelligence (SCAAI), as a Research Intern. Her research interests include the field of artificial intelligence, deep learning, and data science domain, including natural language processing, misinformation detection, multimodal deep learning, credibility analysis studies, explainable AI, and health domain.



POOJA KAMAT received the M.Tech. degree from Mumbai University. She is currently pursuing the Ph.D. degree in the domain of predictive maintenance. She currently works as an Assistant Professor with the Department of Computer Science Engineering and Information Technology, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra, India. She has teaching experience of 12 years and has guided many UG and PG students in the domain of artificial intelligence and machine learning. Her research interests include predictive analytics and its application in the domain of manufacturing, NLP, and healthcare. She has authored many international/national journals and conference publications. According to Google Scholar, she has more than 100 citations, with an H-index of 5 and an i10-index of 5.

SHRUTI PATIL received the M.Tech. degree in computer science and the Ph.D. degree in the domain of data privacy from Pune University. She has been an industry professional in the past, and currently associated with the Symbiosis Institute of Technology, as a Professor, and with SCAAI, as a Research Associate, Pune, Maharashtra, India. She has three years of industry experience and ten years of academic experience. She has expertise in applying innovative technology solutions to real world problems. She is currently working in the application domains of healthcare, sentiment analysis, emotion detection, and machine simulation via which she is also guiding several UG, PG, and Ph.D. students as a domain expert. She has published more than 30 research articles in reputed international conferences and Scopus/Web of Science indexed journals, books with more than 100 citations. Her research areas include applied artificial intelligence, natural language processing, acoustic AI, adversarial machine learning, data privacy, digital twin applications, GANs, multimodal data analysis.

KETAN KOTECHA has expertise and experience of cutting-edge research and projects in AI and deep learning for last 25 years. He has published widely in several excellent peer-reviewed journals on various topics ranging from education policies, teaching-learning practices, and AI for all. He is also a team member for the nationwide initiative on "AI and Deep Learning Skilling and Research" named Leadingindia.ai initiative sponsored by the Royal Academy of Engineering, the U.K. under Newton Bhabha Fund. He currently heads the Symbiosis Centre for Applied Artificial Intelligence (SCAAI). He is considered a foremost expert in AI and aligned technologies. In addition, with his vast and varied experience in administrative roles, he has pioneered education technology. Previously, he has worked as an Administrator at Parul University and Nirma University and has several achievements in these roles to his credit.