FACT - Fine grained Assessment of web page CredibiliTy

by

Shriyansh Agrawal, LALIT Mohan Mohan, Y.Raghu Babu Reddy

in

TENCON 2019 Technology, Knowledge and Society

Kochi, Kerala

Report No: IIIT/TR/2019/-1

Centre for Software Engineering Research Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
October 2019

FACT - Fine grained Assessment of web page CredibiliTy

Shriyansh Agrawal
Software Engineering Research Centre
IIIT Hyderabad
Hyderabad, India
shriyansh.agrawal@research.iiit.ac.in

Lalit Mohan Sanagavarapu
Software Engineering Research Centre
IIIT Hyderabad
Hyderabad, India
lalit.mohan@research.iiit.ac.in

YR Reddy
Software Engineering Research Center
IIIT Hyderabad
Hyderabad, India
raghu.reddy@iiit.ac.in

***Abstract*—With more than a trillion web pages, there is a plethora of content available for consumption. Search Engine queries invariably lead to overwhelming information, parts of it relevant and some others irrelevant. Often the information provided can be conflicting, ambiguous, and inconsistent contributing to the loss of credibility of the content. In the past, researchers have proposed approaches for credibility assessment and enumerated factors influencing the credibility of web pages. In this work, we detailed a $WEBCred$ framework for automated genre-aware credibility assessment of web pages. We developed a tool based on the proposed framework to extract web page features instances and identify genre a web page belongs to while assessing it's Genre Credibility Score (GCS). We validated our approach on 'Information Security' dataset of 8,550 URLs with 171 features across 7 genres. The supervised learning algorithm, Gradient Boosted Decision Tree classified genres with 88.75% testing accuracy over 10 fold cross-validation, an improvement over the current benchmark. We also examined our approach on 'Health' domain web pages and had comparable results. The calculated GCS correlated 69% with crowdsourced Web Of Trust (WOT) score and 13% with algorithm based Alexa ranking across 5 Information security groups. This variance in correlation states that our GCS approach aligns with human way (WOT) as compared to algorithmic way (Alexa) of web assessment in both the experiments.**

***Index Terms*—Web Page, Quality, Ranking, Credibility, Genre, Automated, Framework , Information Security.**

I. Introduction

With the advancement in internet bandwidth, smartphone access and web technologies, creation, delivery and consumption of web content have increased many folds over the past decade. Web has become a primary source for disseminating and accessing information to cater people's content requirement. However, with more than a billion websites present over web constituting to trillion web pages used by 4+ Billion internet users[1], and thousands more getting added on a daily basis, most of the time, users are bombarded with a plethora of information for a simple query. Moreover, this information can be misleading and incorrect, which can have serious consequences for people who increasingly rely on web sources for information such as security, health, academia etc. This favours the use of web search engines, which harvest ranking algorithms to determine the rank of returned search results based on specific preferences. These set of preferences and varied approaches together rank the set of collection distinctively. People often use two divergent terminologies- credibility and popularity, interchangeably. In this modern era of Web 2.0, freedom to content generation and interpolation for end users traditionally leads to a viral spread of misleading and non-credible content, which becomes lethal for domains like security, politics etc. Nevertheless, if the quality of the web page from a source is perceived to be not credible, users should abandon it and explore other sites to meet their requirements [1]. For a domain expert, content relevance, information source, evolution of content and other fine-grained features of credibility decide web page usage [2]. While past researches [3] [4] [5] suggests that novice users, primarily rely on the 'look and feel' of a site. Josang et al. With the variety of content available over the web, in our prior study [6], we propose that the importance of each credibility parameter varies with the genre of a web page. For example (i) the impact of advertisement pop-ups on shopping web pages differs from that on help web pages; (ii) an announcement/notice on portrayal web page has significant impact then on a help web page; (iii) Conventionally, the last modification time has difference importance on news pages vs download pages; (iv) the impact of knowledge base, i.e. text2image ratio of web page differs for news and link-collection web pages. Therefore, to enhance the usefulness of search results, we now focus on genre-aware credibility assessment of a web page.

Genre is an evolving characteristic of communication over any mode (paper, digital or artistic), which is socially constructed, recognised and accepted. Related content/information (document or web page) can be presented in various forms (genre), but the one which are easily identifiable by humans is recognised the most [7]. For instance (i) a letter and a telegram may share the same crucial information despite their form/style is entirely different and thus have individual recognition; (ii) announcements are apt to publish over news web pages and not on help web pages; (iii) briefing of a product is apt on shopping web pages rather than on news web pages. To

[1] http://www.internetlivestats.com

design an automated genre-aware credibility assessment approach, we need to automate the mechanism of genre identification. In the past, researchers [8] [9] have used web genre classification for helping site developers to present their information in a novel way; identification of specific content over the web; for focused crawling etc. Based on our study, there is no prior work on genre classification of web pages for credibility assessment, a fine grained credibility assessment of web page. In 2000, Crowston et al. [10] recognized 48 different genres in web content. Later in 2004, Meyer et al. [11] listed eight generic web genres, namely -{Help, Article, Discussion, Shop, Portrayals of companies and institutions, Private portrayal, Link collection and Downloads}. Based on the reviewed literature on credibility assessment and genre classification, we focus on the following objectives of our current research :

- to focus on genre-aware credibility assessment of web page
- to focus on an approach which identify features and automate web page genre classification
- to propose a process of automated credibility assessment for any generic web page

In our previous work [6], we have proposed an overview of a framework (called $WEBCred$) and developed an elementary tool (based on $WEBCred$) to assess credible score. The prior study focused on credibility assessment from a static point of view with web as a single genre. In this paper, we detail our framework and tool with the extension of automated genre classification approach for multiple genres. This extended tool now automatically assess the Genre Credibility Score (GCS) of a given web page without any human effort unlike before.

As there are more than billion websites with content from various domains, presenting a credibility score for all sites is beyond the scope of our current study. In our prior work [6], we restricted the evaluation of our semi-automated approach to 'Information Security' domain web pages. In the current study, we validate our fine grained automated credibility assessment to 'Information Security' and 'Health' domain web pages. The remainder of the paper is organised as follows: Section 2, we review state of the art on credibility assessment and genre classification of web pages ; Section 3, describes our proposed approach for automation; Section 4, we demonstrate a critical analysis of our results and we present the conclusion and provide an overview of the future work in Section 5.

II. State of the Art

Multiple approaches have been used in the past to counter the problem of credibility assessment and genre classification. The following sub-sections present the past research that we studied before proposing our approach on fine grained genre based credibility assessment of web pages.

A. Credibility

Lack of credibility leads to disinformation and distrust. The importance of credibility on the information source was explained in the seminal work of Hovland et al. [12]. Prominence-Interpretation Theory of Fogg et al. [13] posits that the impact of a web element has on perceived credibility is a product of its prominence (how likely it is to be noticed) and interpretation (what value or meaning people assign to that element). In the early 2000s, researchers [3] [4] [5] have studied and analyzed various features for credibility assessment of the web page adhered to various communities. With the boom of social media and the emergence of web 2.0, the internet has become the primary source of disseminating information, which leads to the tsunami of web pages containing both- credible and fake information. Since then the research focus has been translated to the automation aspects of credibility assessment that led to the development of frameworks [14] [15] and web graph based solutions such as Alexa Ranking[2], PageRank [16]. Later, leveraging crowdsourcing, adaptive solutions were introduced, such as WOT[3], which encourage users to mark the credible ratings of the integrated web pages actively.

The available work on frameworks and APIs do not provide an approach for assessment of generic web pages with possibility of user intervention. The need for user intervention on relevance, results and credibility is widely discussed including in Karen Sparck Jone's speech in acceptance of ACM SIGIR Gerard Salton Award [17]. Various features of web credibility assessment were explained in our prior work [6] which are inline with cognitive heuristics of human judgement for assessment of web pages as explained by Miriam et al. [18]. In our prior user study, we had also validated that the presence or absence of individual web page features, in a particular genre instance, can increase or decrease the credibility of the page. Our previous work, required human effort and extensive validation for a given set of web pages.

B. Genre

Yates et al. [19] suggested that dissemination of information and communication over any new medium will see both emergence and adoption of genres from an existing medium like an article, news (adapted from paper) and Downloads (emerged from Web). The explosion of cheaper smartphones and inexpensive data strengthened the adoption of web pages over paper documents. This motivated Crowston et al. [10] to study the application of genres available in the paper documents to the web. They proposed definitions of 48 different web genres which they further used to classify web pages manually. Meyer et al. [11] identified a list of eight orthogonal and widely used web genres {Help, Article, Discussion, Shop, Portrayal (non-private), Portrayal (private), Link collection, Downloads}. We adopt these genres and automate its identification in our proposed credibility assessment of the web page. Researchers [20] [7] [21] also proposed that

[2]https://www.alexa.com
[3]https://www.mywot.com

dissolution and evolution of genres are socially and computationally demanding. Rehm et al. [22] established a relationship between HTML and web genres to propose a web genre hierarchy for personal homepages of academia. They extended this it from their past attempt [23] to automate the identification of single web genre- 'personal home pages'.

To automate web genre identification, Roussinov et al. [24] and Rehm et al. [22] ignored the evolutionary aspects and looked at it from a static point of view. The work of researchers [24] [11] [22] suggests that web genre is a combination of <content, form, functionality>of web page. Content is the given data on a web page in varied forms (text, image, video etc.). Form or style is the way content presented to the user (HTML properties), and functionality is an evolutionary aspect of the web. Researchers [7] [25] implied the urgency of multi-label genre classification. They proposed that genre classification should respect label from multiple perspectives. Lim et al. [26] made the first attempt to perform a supervised learning approach that uses content and form of a web page to classify a single genre with 75% accuracy. Santini [21] reported an accuracy of about 70% for multi-label classification. Jebari et al. [27] provided an accuracy of about 80% accuracy for multi-label classification with 100+ features which makes the computation expensive. Based on our study, 80% is the current benchmark for multi-label genre classification of web pages. Researchers focused to reduce the number of features and make the process faster with increased accuracy.

III. Approach

In our previous work [6], we had proposed an overview of a framework (called $WEBCred$), which consists of 12 identified features of a web page and detailed their methodology of automated extraction and normalization. We also investigated the individual importance and correlation between these 12 features across genre to establish a scoring mechanism for credibility assessment. In our proposed approach, we extend our features list to automate the identification of a web page genre using supervised learning without any human aid. As a validation of our approach, we applied our approach on two different domains. First, on Information security domain web pages ($P1$), obtained from 157,000 information security web page corpus [28]. Second, on 21,700 Health domain web pages ($P2$), obtained from DMOZ (Open Directory Project, 2016)[4]. $WEBCred$ is flexible and facilitates an increase or decrease in the number of possible genres for a particular domain. Overall, the number of genres of web pages in any domain is potentially an incomplete list. Therefore, we scoped our classification to 7 genres - Article, Help, Shop, Portrayals of companies and institutions (Public Portrayal), Discussion, Link collection and Downloads from the list proposed by Meyer et al. [11] for the sample URLs. We used the web pages from $P1$ as basis of this experimentation. While the web pages from $P2$ were used to validate other domain applicability of our proposed approach. The details of $WEBCred$ framework and feature selection for genre classification is given below-

[4]https://github.com/ALSAREM/dmozNeo4j

A. WEBCred

To identify the most credible web page in a genre of a given corpus, we use/extend $WEBCred$. This framework provides an automated approach of credibility assessment of web pages. The proposed schema of faceted classification by Crowston et al. [7] profoundly inspired the design of $WEBCred$. It accommodates various theoretical structures like - classification, crawling, parsing, etc. and keeps them independent from each other to entertain further extensibility. To provide robustness to tool and survival in the long run, it provides flexibility for user-interruption to afford new features (of any type) at any time with no requirement of complete knowledge.

The process diagram for $WEBCred$ is shown in figure 1. The user enters the URL and waits for $WEBCred$ to

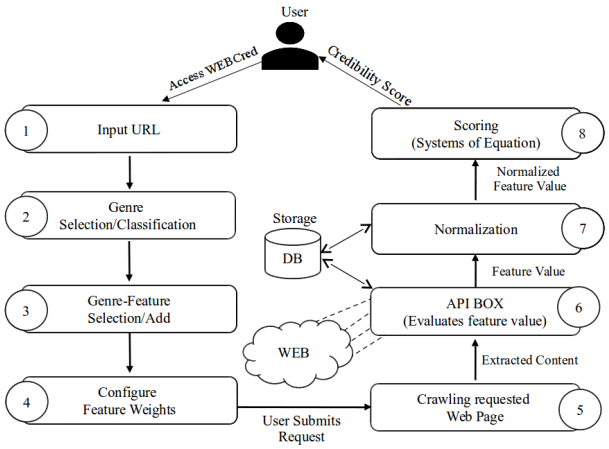

Fig. 1. Process flow of WEBCred [6]

suggest possible genre (Step 1). The genre classification is performed by a supervised learning model that is pre-trained on a labelled dataset. A user may agree with the suggested genre or may select a genre or can define a new one if required (Step 2). For the selected genre, the relevant features are displayed along with their pre-calculated weights (Step 3). However, if required, the user can alter feature weights (Step 4). If the prior assessment of given URL does not exist in the persistent storage (DB) or if the page is modified (i.e. last modified DateTime of the page is altered), then further steps are executed otherwise a credibility score is retrieved from DB. Content of web page is crawled in Step 5 and features values of crawled content is extracted by an utility called $APIBox$ at Step 6. The extracted values are then normalized in a range of (-1,0 and 1) at Step 7, which is then used to calculate Genre Credibility Score (GCS) at Step 8. For the validation of our proposed approach, we also developed a tool based on our proposed framework, which is

Feature name	Attributes	Extraction API/Methodology
Keywords	Top 10 keywords of web page	Words of web page are ordered based on their Tf-idf scores[5]
Sentiment	Count of Positive, Negative and Neutral Sentences	Using sentiment annotator of Stanford's CoreNLP library [6]

TABLE I
LEXICAL FEATURES OF A WEB PAGE

Feature name	Attributes	Extraction API/Methodology
POS tag	Count of individual POS tag in page's text	Using POS taggers of Stanford's CoreNLP library [6]
Count of Symbols	Currency, Date, Scientific Units, Abbreviation, Shop keywords (sell, buy, purchase, cart), Help Special Keywords (FAQ, help, support)	Matching words with self composed regular expressions
Count of Contact info	Email, Phone, Address, Names, Social Network info and Matching words with self composed regular expressions	
Count of Text Tokens	Sentences, Words, Characters, Digits, Individual punctuation marks	Using Text tokeniser of Stanford's CoreNLP library [6]
Count of Misspell	Spelling errors in text	Counting words which are not present in dictionary of NLTK[7] library

TABLE II
TOKEN FEATURES OF A WEB PAGE

deployed at [29] and it's source code is available on Github at [30]. The major steps of $WEBCred$ framework are detailed below.

B. Features Selection and Extraction

Web pages evolved from containing static content to machine-generated dynamic (scripts based) content. The purpose of a web page, i.e. information dissemination, is still intact. The characteristics of web elements are referred to as features in this paper and are used to classify genre and assess credibility. Based on available literature [24] [11] [22], we classified the primary elements of web page into 3 sub-categories namely <Content, Form, Functionality> for ease of understanding.

<Content> is the given text on page and is widely used for genre classification of paper documents [11]. We sub-categorise it into Lexical and Token information which together carries information from various dimensions of the text and helps in analysing their individual importance. <Form> or style is the way content is presented to the user. For web pages, it can be analyzed based on its URL and HTML properties [26]. <Functionality> of a web page is evolving with the web over time. Since the emergence of web pages, they have been evolved from static (plain) content to user-interactive dynamic (scripts based) content. Based on prior research [6], surface features of a web page are given the primary focus in this category. Surface features of a web page tend to be dynamic and are not recognised in any other medium. In the following subsections, we detail each of these sub-categories and their respective features along with the methodology used for their inspection/extraction.

1) Lexical and Token: Lexical information is a widely used parameter for text processing models [31]. We considered keywords and sentiment as valuable markers to a moderate web page. Keywords describe the contents of a web document based on a given text corpus. Sentiment defines the attitude of content writers for the topic or the overall contextual polarity or emotional reaction to the web document. The token analysis includes information on individual sentences, words and characters in the form of frequency (count) of - POS (parts of speech) tags; Symbols; connection info and individual text tokens. Sentiment annotators, Text tokeniser and POS taggers of Stanford's CoreNLP library[6] are used to extract related sentiments, token information and POS tag out of web text. We compose regular expressions to obtain frequency of individual matched symbols and contact info. NLTK[7] library is used to check misspells on web page. Table I and II lists all features along with their extraction methodology used in this study.

2) URL: Uniform Resource Locator (URL) defines the location of a web resource on web. A typical URL[8] consists of <protocol> (http); <hostname> (www.example.com) that contains <Generic Top-level Domain (gTLD)> (com); optional <port> (8080) separated by ':' ; optional <path> (/../); <filename> (index.html) and optional <query> (q=search) separated by '?'.

<protocol>, <port> and <hostname> , are the elementary properties of any URL and therefore ignored in this study. <query> is also ignored in this study as it varies depending on the query sent to a database. Presence of specific lexical terms in URL often gives a vague presence of a specific genre. Inspired from the work of Lim et al. [26], we included lexical terms that occur more than three times in URL strings of the training corpus.

We used Python's URLlib module along with self-composed regular expressions to extract feature values as shown in Table III.

3) HTML: HTML tags are the building blocks of web pages. The tags regulate styling of a web page and provide a means to create structured documents. Of all these tags, anchor tag (<a>) with attribute HREF is used for referring to other pages, count of which is commonly used for ranking web pages [16]. The count of brokenlinks (HTTP status code between 400 - 500) and the count of outlinks (referring to other websites) on the webpage are identified. We further inspect frequency (count) of all HTML tags used in a web page. Table IV lists all HTML features along with extraction methodology used in this work.

[6]https://stanfordnlp.github.io/CoreNLP/annotators.html

[7]http://www.nltk.org/

[8]http://www.example.com:8080/../index.html?q=search

Feature name	Attributes	Extraction API/Methodology
Depth of URL	Number of directories included in <path>	Split <path>by "/"
gTLD	Domain area (com, gov, org etc.)	Used Python's urllib module
Document type	extension in <filename> (html, script, doc etc)	Used Python's urllib module
Presence of lexical terms	papers, start, file, gallery, introduction, info, login, search, research, bbs, link, intro, people, profile, video, photo, faq, news, board, detail, list, qna, index, shop, data, view, front, main, company, item, paper, product, read, sell, buy, purchase, support, help, cart	Matching words with self composed regular expressions (True if present else False)

TABLE III
URL FEATURES OF A WEB PAGE

Feature name	Attribute	Extraction API/Methodology
HTML	Count of all HTML tags	Parsing html content using Python's BeautifulSoup module
OutLinks	Count of links which point to outer websites	Reckon all hyperlinks which redirect to a different <hostname>
BrokenLinks	Count of error links	Hyperlinks which return with status code greater than 300

TABLE IV
HTML FEATURES OF A WEB PAGE

Feature name	Attribute	Extraction API/Methodology
Advert-isements	Count of banner advertisements and unwanted frames	Using ReGex filter of Easylist[9], which removes most adverts from webpages
Page Load Time	Response time (in sec) of request for a web page	Using Python's request Library
Responsive Design	Whether automatic alteration of a web page is rendered based on the screen resolution of device being used to view it	Using Google's FriendlyTest Api[10]
Modified DateTime	Time (in sec) since page is last updated	Either using Archive[11] data or in web page's metadata
InLinks	Count of out of <hostname> links pointing to given webpage	Using Google's search Api[12]
Inter-nationali-zation	Count of language options for website	Count of distinct value in <lang> tag of HTML content
Real World Presence	Presence of certain connection information (email, address, phone, social media presence, members)	ReGex match in header and footer tag of HTML
Text2Image Ratio	ratio of viewport (viewing region) of text and non-text (images, video)	Calculating ratio of viewport (viewing region) of text and non-text (images, video)

TABLE V
SURFACE FEATURES OF A WEB PAGE

4) Surface: Elements of a web page that are dynamically rendered on a screen or adhere to functionality and are user-interactive are considered as surface features [6]. For a novice user, surface features such as fonts, colour, images and other layouts of the web page create the first impression. All such features along with the respective description and extraction methodology used are listed in table V.

Banner or advertisements reduce the credibility of web page content specifically when the advertising is unexpected; Page Load Time (response time) affects the credibility of certain web pages such as help/support pages; Need of responsive in design and web page internationalization (varied language support) with growing usage of internet in developing countries (regional languages) on smaller screens (including smartphones); Modified DateTime provides a measure on the web page freshness; Inlinks has been the traditional way to inspect popularity; Evidence of real World presence in webpage asserts the existence of author/owner; Text2Image ratio suggest availability of knowledge base.

The open APIs used for extracting surface features in a web page are : Easylist[9] service used by popular ad-blockers to identify advertisements in web pages, including unwanted frames and images; Python requests library to check the response time of the web page; Mobile FriendlyTest[10] of Google API service to validate a web page responsive design behaviour; Web Archives[11] APIs to identify last modified DateTime based on the web page metadata; Google API[12] to extract inlinks of the web page; Internationalization is inspected based on the lookup for <lang> tag in HTML content of web page; Presence of certain connection keywords in header and footer of a webpage or top and bottom 20% of HTML body to inspect real-world presence and the Text2Image ratio of a web page based on the ratio of viewport (viewing region) of text and non-text (images, video). These APIs are selected based on the available literature, their wider usage count, ease of code integration and the licensing terms.

5) $APIBOX$: To ease the extraction of feature instance values for crawled web pages, we developed an automated tool, called $APIBOX$ with persistent storage (DB). The $APIBOX$ crawls randomly selected 10% of 157,000 sample 'Information Security' ($P1$) web pages for overall validation of approach. And 50% of 21,700 sample 'Health' domain web pages ($P2$) for cross-domain validation of approach. The crawled content is parsed to extract text and HTML of web page after removal of Javascript and other verbose code. The obtained content is stored in text and HTML corpus of DB and then $APIBox$ initiates threads to obtain the value for all above mentioned features from the crawled web page. Out of all crawled web pages, only 10,429 (S') of $P1$ and 8,143 (V') of $P2$ each were considered for the study as others had HTTP errors during

[9]https://easylist.to/easylist/easylist.txt
[10]https://search.google.com/test/mobile-friendly
[11]https://archive.org/
[12]https://developers.google.com/custom-search/json-api/v1/overview

extraction. A snapshot of a sample set of pages with feature value instances. The source code of $APIBox$ is available at GitHub repository[13] and collected data by $APIBox$ is also available in it's sub-repository[14].

6) Normalization: Features' value provided by $APIBox$ are on different scales for web pages. Also, there may exist some APIs or extraction methodology which may provide enhanced results compared to the ones we have used. Therefore to avoid such discrepancy, we normalized our extracted feature values for all S' and V' web pages. Features such as Internationalization, Misspell and Responsive Design are normalized to 0 or 1 based on it's presence. Modified DateTime value is normalized to 1 if the web page is updated within a month ($< \ 30days$, configurable) otherwise 0. We considered 6 Top Level Domain ($gTLD$) with scores as : 1 - $.gov$, 1 - $.edu$, 0 - $.org$, 0 - $.com$, 0 - $.net$ and -1 to all others. While there is no documented guideline, it is generally acknowledged by the user community that the content in $.gov$ and $.edu$ has more credibility. To normalize all other features' (f_i) value to a measurable scale, the same sample set of S' information security web pages (wp) were used to calculate the mean (μ) and the standard deviation(σ) value of each individual features. Each feature value (f_i) for a given sample is normalized (v_i) to $\{-1, 0, 1\}$ across genres based on Eq.1

$$v_i = \left\{ \begin{array}{cc} \text{-1 ,} & \text{f}_i < \mu \text{ - } \sigma \\ \text{ 0 ,} & \text{f}_i \in [\mu \text{ - } \sigma, \mu + \sigma] \\ \text{+1 ,} & \text{f}_i > \mu + \sigma \end{array} \right\} \quad (1)$$

C. Genre-Classification

To classify the genre of a given URL, we experimented with supervised learning techniques to train a model for active classification. To train the learning model, we conducted a reward based task to label the S' dataset leveraging a crowdsourcing platform.

1) Labels: We conducted an online labeling task on paid crowdsourcing platform provided by CrowdFlower[15] (now known as Figure Eight Inc.) and unpaid online survey[16] which have participation from academia. The dataset of S' URLs was prepared for genre labeling.

Both, the crowdsourcing task and survey had an explanation of the underlying task with examples for ease of understanding. Each URL was labelled by four CrowdFlower workers and once by an academic participant to obtain quality responses. A crowdsource worker could provide a maximum of 200 judgments, where each judgment consists of two test questions (URLs which are pre-labelled by authors) and eight unlabeled questions. Workers can choose among labels- {Help, Article, Discussion, Shop, Portrayals of companies and institutions, Link collection, Downloads Broken Links and Others} for every URL. Each worker is paid 3 cents for labeling a URL, and it takes approximately 1 - 1 1/2 hours for labeling 200 URLs. 200 such test questions were pre-labeled by authors for quality check. Few of the test questions were broken links that were purposefully added to check the seriousness/quality of responses. The worker was credited with rewards, and their response was taken into account only if the worker maintained accuracy of 80% consistently with test questions throughout their judgments.

CrowdFlower, as a platform had a mechanism to filter out unreliable, inexperienced and unregistered workers. Through the platform, we obtained workers with prior experience in labeling tasks to get quality output. 5,787 workers from 25 different countries labelled S' URLs. 4,280 URLs of the S' dataset were also labelled by 37 students from academia.

Among 5 (in some cases 4) labels for each URL, the one with highest confidence score[17] was used for further synthesis. After labeling by workers, URL with confidence score less than 1 for the classified label was discarded from the final dataset. As a result, 8,550 (S) URL-Label sets were considered to train models for genre classification, as detailed below.

2) Dataset: A training dataset consisting of 8,550 URLs with labels and feature value is prepared. Labelled dataset is available on GithHub[18]. Few of the features have dictionary type values which are not easy to parse by learning models. Therefore, we flattened all dictionary keys into individual features which results in the count of total 688 features. Further, each individual string value of a feature has been assigned a categorical value to feed to the training model. Most of the samples were partially filled due to presence of certain feature, so we assign all empty cells with value "0". To reduce redundancy, we removed features from dataset that have variance value less than "0.1" as they contain very less information pertaining to classification. 517 such features were removed which finally results to a dataset of $8,550 \ X \ 171$ size. Computing 171 feature values in run time, requires more extended time and power for active classification. To reduce features, we created two copies ($D1$ and $D2$) of the dataset and applied feature selection techniques. Features are reduced using two different techniques - Anova for $D1$ and Mutual Information Gain (MI) for $D2$.

Stable and consistent threshold value of Anova and MI score for filtering ensures high testing accuracy with lesser number of features. The figure 2 & 3 shows the Anova and MI score of 171 features. We performed the experiment over a range of Anova and MI scores to select features. The experimental Anova score ranges from 0.5 to 7.5 with an interval of 0.5 and MI score ranges from 0.105 to 0.17 with an interval of 0.005. We now have 13 datasets of $D1$ and 12 datasets of $D2$, each having number of features depending upon the threshold value. To have equal distribution of genre

[13]https://tinyurl.com/ApiBox-SourceCode
[14]https://tinyurl.com/webcred-featureData
[15]https://make.figure-eight.com
[16]https://form.jotform.me/82411827578464
[17]https://tinyurl.com/FEConfScore
[18]https://tinyurl.com/webcred-genreData

classes, we performed oversampling on all of these 25 (12 + 13) datasets before training the supervised learning models.

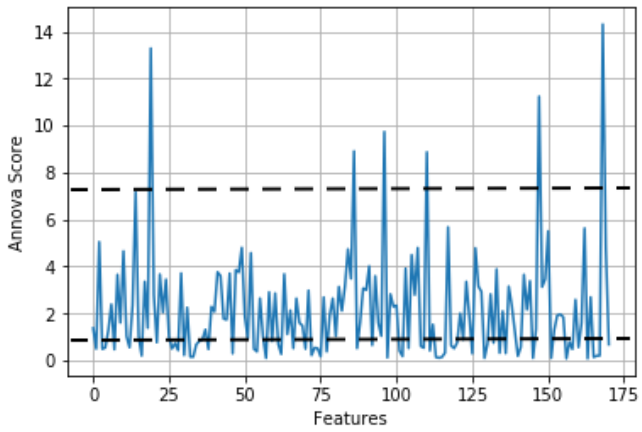

Fig. 2. Anova Score of Features

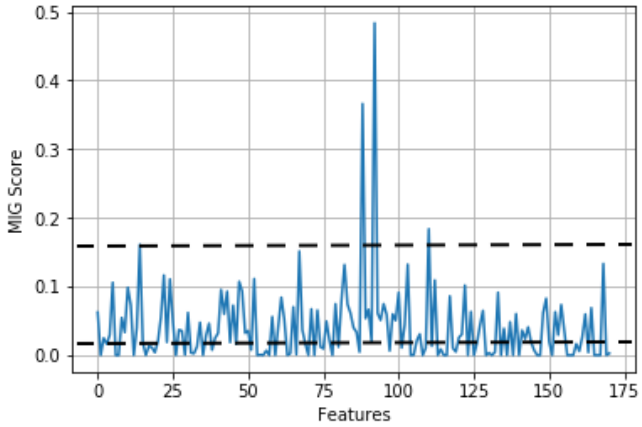

Fig. 3. MIG Score of Features

3) Experimental Settings: To classify Genre of web page based on the identified features, we used our S dataset with 4 classical supervised learning models - Multi-Layer Perceptron based Neural Networks (network based learning) with two hidden layers of nodes 128 and 32 respectively; Support Vector Machine (that applies kernel trick to handle higher dimensions) with balanced class weight; Gradient Boosted Decision Trees (based on probability) with max depth set to 3 ; Logistic Regression model (based on linear combination of independent features)

D. Scoring

A discussed earlier Web Page can be divided into 3 key elements namely- {content, form and functionality }. Each of these elements has numerous sub-elements coined as features of web page in this study. We established in our previous work [6], that twelve of these features namely :{ Advertisements, Internet domain ($gTLD$), Real World Presence, Misspell, Text2Image Ratio, Modification DateTime, [In, Out, Broken] Links, Internationalization, PageLoadTime, Responsive Design } are orthogonal in nature and are acceptable for credibility assessment. Orthogonality of these features ensures that there is null or negligible inter-dependency amongst the features and therefore a linear equation can be established for assessment. Hence an experimented linear equation (Eq. 2) was formulated where it's coefficient (w) signifies importance of each individual feature (f_i) for classified Genre (g) and (v) represents normalized value of each feature (f_i) of web page (p). We call this credibility value as Genre Credibility Score (GCS) of Web Page in this study.

Genre	Count
Article	6390
Help	230
Shop	310
Public Portrayal	780
Discussion	270
Link Collection	320
Downloads	250

TABLE VI
COMPOSITION OF INFORMATION SECURITY DATASET

$$GCS_p = \sum_{i=1}^{n} (w_g^{f_i} * v_p^{f_i}) \tag{2}$$

In our previous work [6], we have also established the phenomenon, that the importance of individual features varies w.r.t genre of a web page. For example, several pop-ups and advertisements on shopping web pages may not bother users as much as on article web pages. We further surveyed to calculate individual importance of each feature across different genre as shown in figure 4 from prior study. This calculated weightage of each feature-genre pair is then included as default in scoring.

Category	Articles	Help	Shop	Portrayal -Org
Domain	0.079	0.059	0.069	0.111
Advertisements	0.053	0.059	0.138	0.083
Text2Body Ratio	0.105	0.059	0.103	0.056
In Links	0.105	0.059	0.069	0.083
Misspell	0.079	0.059	0.069	0.111
PageLoad Time	0.079	0.059	0.138	0.056
Broken Links	0.053	0.059	0.103	0.111
Presence (Contact Info)	0.053	0.059	0.103	0.111
Responsive Design	0.105	0.059	0.069	0.083
Last Modified Date	0.105	0.059	0.069	0.083
Internationalization	0.105	0.176	0.034	0.056
Outlinks	0.079	0.235	0.034	0.056

Fig. 4. Feature Weightage for Genres [6]

IV. RESULTS AND ANALYSIS

A. Classification Results

We used the S (8,550 URLs with 171 features) dataset to train supervised learning models– Multi-layered perceptron based Neural Network (MLP), Support Vector Machines (SVM), Gradient Boosted Decision Trees (GBDT) and

Logistic Regression (LR) for web page genre identification. The Table VI represents initial composition of our dataset, which is later balanced using oversampling technique to contain a minimal of 450 samples of each genre.

Figures 5 & 6 represents testing accuracy of the four models vs varied feature count (features selection based on their Anova Score and MI Score). The models were 10-fold cross-validated to reduce variance and bias. As evident from

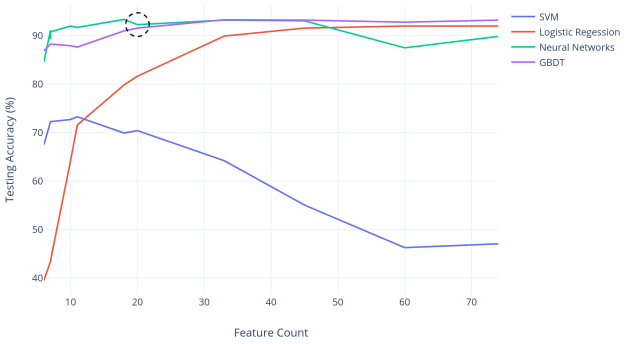

Fig. 5. Anova selected Features vs Testing Accuracy

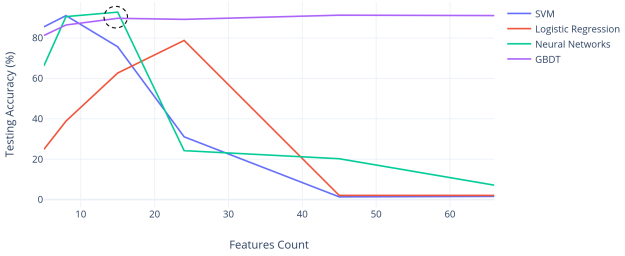

Fig. 6. MI selected Features Vs Testing Accuracy

figure 5 & 6, at the encircled portions, the testing accuracy has been stabilized w.r.t features count. Also, it is observed that all 4 models performed better (in terms of testing accuracy) with Anova filtering for a varied range of scores in comparison to MI filtering. Therefore, we selected Anova filtering over MI. Further analysis of Figure 5 at encircle, establish that Gradient Boosted Decision Trees ($GBDT$) with feature count -20 (Anova threshold value - 4.5) had higher testing accuracy alongside low count of features. Evidently, we selected $GBDT$ with Anova filtering as our final classification model for Genre identification.

Table VII present all those twenty (out of 171) identified features, grouped according to their sub-category. Few of these features are also a part of credibility assessment, which are– ads, outlinks (represented by a html tag) and domain, while others purely account for form and content of genre of

Feature Name/Category	Attribute
POS Tags	SYM, RP, WDT, VBZ, CC, CD, JJR, MD, RB,
HTML Tags	img, form, meta, a area, textarea, td, map
URL	Depth, Domain
Surface	Advertisement

TABLE VII
CATEGORISED GROUP OF 20 REDUCED FEATURES FOR GENRE CLASSIFICATION

Class Name	Precision	Recall	F1-Score
Article	0.81	0.77	0.79
Help	0.97	1	0.99
Shop	0.78	0.68	0.73
Public Portrayal	0.98	0.96	0.97
Discussion	0.94	1	0.97
Link Collection	0.87	0.8	0.84
Downloads	0.82	1	0.9

TABLE VIII
CLASSIFICATION OF $GBDT$ WITH ANOVA VALUE-4.5

web page. One can refer to description of individual POS tag here[19]. Moreover, computation of these 20 feature instances is not to complex and can be computed in run time, which serves the purpose of feature reduction.

The average testing accuracy for this selected model is 88.75% over 10-Fold Cross Validation. Table IX shows the percentage of correctly classified pages on the diagonal and summarizes the percentage of misclassified pages with respect to other genres. Table VIII represents Precision value, Recall value and F1-Score for all 7 genres. High precision value over recall allows to keep a balance between them to maximize the value of F1-Score.

However, these models so far have only been trained and tested with Information Security web pages ($P1$). So we further tested our trained model ($GBDT$) with the selected feature set over Health domain web pages ($P2$). For the validation purpose, we picked 400 URLs (from V', 8143 web pages) on the basis of the availability of their alexa, wot and pagerank data; and one of the authors labelled them with respective genres. Moreover, a co-author cross-examined those labels to maintain consistency. Table X represents the initial composition of 400 labelled URLs with respective genres. We were not able to gather URLs of other genres; however, that is not the key concern here, as a collection of all possible genres is beyond the scope of this study. The labelled data of Health domain is available at Github repo[20].

Our trained model correctly identified genre of 329 URLs, i.e. with 82.26% accuracy. Table XI shows the percentage of correctly classified pages on the diagonal and summarizes the percentage of misclassified pages with respect to other genres. As can be observed $GBDT$ model has been consistent across genres of unseen domain web pages as well. Hence, we rely on $GBDT$, an ensemble learning algorithm that optimizes the loss function while providing higher accuracy, thus, the model can be extended for additional datasets and features.

[19]http://tinyurl.com/havrq57
[20]https://tinyurl.com/genre-Health

	Article	Help	Shop	Public Portrayal	Discussion	Link Collection	Downloads	Total
Article	76.92%	2.56%	2.56%	10.26%	5.13%	0.00%	2.56%	100%
Help	0%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100%
Shop	0.00%	0.00%	95.56%	0.00%	0.00%	4.44%	0.00%	100%
Public Portrayal	17.07%	0.00%	0.00%	68.29%	0.00%	7.32%	7.32%	100%
Discussion	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	100%
Link Collection	0.00%	0.00%	0.00%	9.76%	2.44%	80.49%	7.32%	100%
Downloads	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	100%

TABLE IX
TEN-FOLD CROSS-VALIDATED CONFUSION MATRIX OF $GBDT$

Genre	Count
Article	112
Public Portrayal	214
Discussion	22
Link Collection	52

TABLE X
COMPOSITION OF HEALTH DATASET

Jebari et al. [27] provided an accuracy of about 80% for multi-label classification but with 100s of features which make the process infeasible on run time. To the best of our knowledge based on available literature 80% is the current benchmark of multi-label genre classification of web pages. Therefore, it is now evident that our approach with accuracy 88.75% crossed current benchmark with lesser number of required features set.

B. Validation of Our Approach

For validation of overall approach, we examined it independently with 'Information Security' and 'Health' domain web pages. First, 40 new URLs from 5 individual Information Security Groups have been collected from past research [28]. Feature values of all these 200 ($40X5$) URLs are extracted and normalized by $APIBOX$. And genre of every URL is identified by our trained model ($GBDT$ with Anova reduction) to get applicable feature weightages for scoring. Second, we selected 200 URLs of Health domain , which were classified by our trained model and normalized feature instances of every URL were collected by $APIBOX$. The $WEBCred$ calculated GCS score of all these URLs based on the identified genre and normalized feature values. These calculated GCS scores are then compared (using cosine similarity) with Alexa Ranking (popularity based, widely used by search engines) and Web Of Trust (WOT) Ranking (crowdsourced/reviews based) for each URL.

Table XII shows the correlation of GCS Vs Alexa, GCS Vs WOT and Alexa Vs WOT of selected 5 Information Security Groups. For 'Information Security' web pages, GCS correlates 13.52% with the Alexa ranking and 69.48% with WOT ranking. Between Alexa and WOT, the correlation is 12.31%. High variance in correlation states that all approaches work on different orientation. On the other hand, for selected 200 Health domain the GCS correlates 15.17% with the Alexa ranking and 59.9% with WOT ranking. Between Alexa and WOT, the correlation is 23.83%.

Disparity in correlation results of two individual domain-experimentation is advocated by– variation in correlation score of WOT and Alexa ranking between the two domains (12.32% for Security while 23.83% for Health). Which imply that the two domains does not share similar feature instance importance and eventually end up having distinct correlation scores. $WEBCred$ encompasses genre as a selection criteria to get importance of individual features that are not being used by other algorithms such as PageRank, as the credibility of the web page is not their primary motive but popularity. It can be observed that GCS score has better correlation with WOT in both domains confirming that our approach aligns with the human way of web page assessment.

V. CONCLUSIONS AND FUTURE WORK

We designed $WEBCred$ framework for genre-aware automated assessment of credibility of a web page. The framework automates– feature extraction and normalization; $GBDT$ supervised trained model for genre classification; and scoring mechanism to calculate Genre Credibility Score (GCS) of a web page. It also provides possibility for user intervention to add/reomve features and genres as per requirements. Our genre identification approach overpasses the current benchmark of 80% classification accuracy with lesser number of required features sets. Further, we developed an Open Source tool based on $WEBCred$ to validate our approach. We validated our approach with 'Information Security' web pages and as a further validation across 'Health' domain web pages. The GCS calculated by our tool correlated 69% with WOT Score and 13% with Alexa ranking across 5 security groups. And similar correlation results for 'Health' domain web pages, which advocates that our approach can be extended to additional web domains as well. High variance in correlation states that all approaches work on different orientation, where our approach is more inline with human way (WOT) of judgment.

The web is evolving with time, and there can be many more related features in the future, which is why it is quite difficult to create a complete, reliable set of features of web page, but completeness is not the key concern of our study as we anticipate the possibility of adding many more features in future if required. In future, we propose to extend our work in the following ways:

	Public Portrayal	Article	Link Collection	Discussion	Total
Public Portrayal	84.62%	11.54%	3.85%	0%	100%
Article	4.55%	81.82%	4.55%	9.09%	100%
Link collection	0%	6.67%	80%	13.33%	100%
Discussion	0%	4.35%	13.04%	82.61%	100%

TABLE XI
TEN-FOLD CROSS-VALIDATED CONFUSION MATRIX OF $GBDT$ OVER HEALTH DOMAIN WEB PAGES

Group	Alexa Vs WOT	GCS Vs Alexa	GCS Vs WOT
Attacks	11.43%	20.06%	72.67%
Cloud Computing	11.45%	22.45%	60.98%
Endpoint	4.35%	8.72%	70.64%
Network	19.14%	12.65%	62.83%
Cyber	15.16%	3.72%	80.26%
Average	12.31%	13.52%	69.48%

TABLE XII
CORRELATION OF GCS WITH ALEXA AND WOT

- Validate our approach with more URLs in information security and other domains such as education. Also, we plan to study the same for web pages that may fall into multiple domains and genres.
- Validate the ranked web pages by domain experts for agreement on the genre based approach.
- Develop a browser plugin for $WEBCred$ that can be used online for credibility assessment of search engine results.

REFERENCES

[1] A. Jøsang, C. Keser, and T. Dimitrakos, "Can We Manage Trust?" in *Proceedings of the Third International Conference on Trust Management*, 2005.

[2] R. Ahmad, A. Komlodi, J. Wang, and K. Hercegfi, "The Impact of User Experience Levels on Web Credibility Judgments," in *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, 2010.

[3] B. J. Fogg and H. Tseng, "The Elements of Computer Credibility," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1999.

[4] C. N. Wathen and J. Burkell, "Believe it or not: Factors Influencing Credibility on the Web," *Journal of the American Society for Information Science and Technology*, 2002.

[5] J. Lazar, G. Meiselwitz, and J. Feng, "Understanding Web Credibility: A Synthesis of the Research Literature," *Foundations and Trends in HumanComputer Interaction*, vol. 1, no. 2, 2007.

[6] S. Agrawal, S. L. Mohan, and Y. R. Reddy, "Automated credibility assessment of web page based on genre," in *Big Data Analytics*, 2018.

[7] K. Crowston and B. Kwasnik, "A framework for creating a facetted classification for genres: addressing issues of multidimensionality," *37th Annual Hawaii International Conference on System Sciences*, 2004.

[8] G. T. De Assis, A. H. Laender, M. A. Gonçalves, and A. S. da Silva, "A genre-aware approach to focused crawling," *World Wide Web*, 2009.

[9] I. Pollach, "Electronic word of mouth: A genre analysis of product reviews on consumer opinion Web sites," *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2006.

[10] K. Crowston, "Reproduced and emergent genres of communication on the World-Wide Web," in *Proceedings of the Thirtieth Hawaii International Conference on System Sciences*, 1997.

[11] S. Meyer Zu Eissen and B. Stein, "Genre Classification of Web Pages," *Proceedings of KI04 27th German Conference on Artificial Intelligence*, 2004.

[12] C. I. Hovland and W. Weiss, "The Influence of Source Credibility on Communication Effectiveness," *Public Opinion Quarterly*, 1951.

[13] B. J. Fogg, "Prominence-interpretation Theory: Explaining How People Assess Credibility Online," in *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, 2003.

[14] S. S. Sundar, "The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility," *Digital media, youth, and credibility*, 2007.

[15] Y. Yamamoto and K. Tanaka, "Enhancing Credibility Judgment of Web Search Results," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab," Tech. Rep., 1999.

[17] K. S. Jones, "A Look Back and a Look Forward," in *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1988.

[18] M. J. Metzger and A. J. Flanagin, "Credibility and Trust of Information in Online Environments: The Use of Cognitive Heuristics," *Journal of Pragmatics*, 2013.

[19] J. Yates and W. J. Orlikowski, "Genres of Organizational Communication: A Structurational Approach to Studying Communication and Media," *Academy of Management Review*, 1992.

[20] W. J. Orlikowski and J. Yates, "Genre Repertoire: The Structuring of Communicative Practices in Organizations," *Administrative Science Quarterly*, 1994.

[21] M. Santini, R. Power, and R. Evans, "Implementing a Characterization of Genre for Automatic Genre Identification of Web Pages," *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 2006.

[22] G. Rehm, *Hypertext Types and Markup Languages*, 2010.

[23] ——, "Towards automatic Web genre identification: A corpus-based approach in the domain of academia by example of the Academic's Personal Homepage," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2002.

[24] D. Roussinov, K. Crowston, M. Nilan, B. Kwasnik, J. Cai, and X. Liu, "Genre based navigation on the web," in *Proceedings of the Hawaii International Conference on System Sciences*, 2001.

[25] M. Santini, "Characterizing genres of web pages: genre hybridism and individualization," *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2007.

[26] C. S. Lim, K. J. Lee, and G. C. Kim, "Multiple sets of features for automatic genre classification of web documents," *Information Processing and Management*, 2005.

[27] C. Jebari, "Enhanced and combined centroid-based approach for multi-label genre classification of web pages," *International Journal of Metaheuristics*, 2015.

[28] S. L. Mohan, S. Sarangi, Y. R. Reddy, and V. Varma, "Fine Grained Approach for Domain Specific Seed URL Extraction," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE, 2018.

[29] S. Agrawal, S. L. Mohan, and Y. R. Reddy, "Webcred user interface," https://tinyurl.com/WEBCredFramwork, (accessed 4-May-2019).

[30] ——, "WEBCred Source Code," https://tinyurl.com/WebCredFramework, (accessed 4-May-2019).

[31] B. Kessler, G. Numberg, and H. Schütze, "Automatic Detection of Text Genre," in *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, 1997.