Automated Assessment of Website Credibility

Douglas Fraser
UEA Registration: 100189521

December 15, 2016

Abstract

Evaluating the credibility of information available on the Web is important given that anyone can publish anything and reach millions of people. So what are the components of a website that make it credible (or not)? How can these be determined, analysed, and verified? There are many possibly useful factors, not just the veracity of the information. This paper reviews previous research into website credibility and proposes a design for DIOGENES, an end-to-end system that assists users in evaluating websites. As part of the review, high level categories of website characteristics and the different approaches to assessing credibility are discussed. Also analysed is how the type of website and the larger goal in establishing credibility both affect the evaluation process.

1 Introduction

"On the Internet, no one knows you're a dog" (Steiner, 1993) is a famous cartoon that says something significant about the nature of the Web. There are tremendous amounts of information and services available, but how much is actually verifiable or trustworthy? In other words, is the website credible? Given the importance of the Web and the services it provides and facilitates for society, establishing the credibility of a website is crucial. Otherwise, the services the website purports to offer could harm a user in various ways. Examples of such harm can be found in Table 1.

Table 1: Categories of harm and Examples of websites causing harm

Category	Example Website
Financial	A website that imitates a banking website in order to steal credentials
Medical	Websites that promote ineffective or unregulated cancer treatments
Intellectual	Websites that disseminate false information about celebrities or hoaxes
Social	Websites that spread manipulative political news and propaganda
Security	A website that spreads malware to hijack users' computers

These are not theoretical examples, unfortunately. The problem of assessing whether a website can be trusted is a current and pressing one. The number and variety of websites are increasing every year due to the growth of the Internet (just over 1 billion websites as of October 2016 (Netcraft Survey, 2016)). Consequently, straightforward approaches to assessing credibility (e.g. manually maintaining whitelists or blacklists) are impractical. Therefore automated solutions are required, if only to assist humans in the task.

There is much research into credibility available, but as it is a multifaceted concept, there are numerous aspects to the results of these studies. Section 2 discusses the details of the concept of credibility, highlighting important aspects to consider. Section 3 reviews the research, examining and categorizing the different types of website characteristics used in assessments. Section 4 categorizes the different approaches the studies use, examining their strengths and weaknesses. Sections 3 and 4 form the basis for section 5 which discusses how the type of website, the nature of the data, and the particular goal affect the assessment process. Finally, section 6 outlines at a high level a proposed system for automation of credibility assessment. The system is named DIOGENES after the Greek philosopher who searched the world looking for an honest man. Both the user interface for a browser extension and the back-end it communicates with are discussed. Section 7 is a summary of this paper.

2 Components Of Credibility

The examined literature reveals multiple ways of looking at credibility because it is a complex and multifaceted concept. One reason for the complexity is that, besides the objective aspect, there is a subjective aspect to a full assessment of credibility. Therefore, characteristics of the viewer must be taken into consideration. Another complication is that credibility involves evaluating something along multiple dimensions. This is because credibility is comprised of simpler concepts; different scientific disciplines tend to emphasize different facets of credibility. An automated system should account for all these facets in order to be useful. A third complication is that there are different types of credibility associated with an information source. This section attempts to summarize these specific considerations; they are important in developing a system that assists users in evaluating credibility.

Explaining the objective/subjective dichotomy involves examining how credibility is defined by different sources. Fogg and Tseng (1999), in their landmark study of computer system related credibility, define it in terms of trustworthiness and expertise. However, credibility is typically defined in dictionaries using the base concepts of trust and believability (Credibility, nd). So what is the difference between expertise and believability? Expertise is a more objective concept in that it is associated with knowledge, competence, etc. – things that can be objectively assessed. But believability is a subjective concept in that it is associated with a viewer's psychological state of mind, i.e. what they believe. This means the viewer's interpretation can never be discounted – of the information, the veracity of it, or even of the expert responsible for verifying the information. In other words, this means the viewer's psychological processes are as relevant as any objective measures.

As an example of how psychological processes are important, consider the interpretation of the websites in Table 1. The biased descriptions are used by this author to establish which websites are untrustworthy and why. Without those adjectives, the credibility of the example websites would depend even more upon the reader's own interpretation of the description. Thus the reader's assessment of this author's expertise – or truthfulness – can be seen to be a factor. This leads to the conclusion that believability can be thought of as 'perceived expertise' and so is slightly different than objectively based or assessed expertise.

A thorough examination of users' psychological processes in evaluating website credibility is in Fogg et al. (2001), Fogg et al. (2003) and Kakol and Nielek (2015). Users were asked to note what aspects of a website were the most prominent for them in assessing credibility and to explain their reasoning. The results show the top consideration was the quality of the graphic design of the site. How well the information was structured was the second. Surprisingly, the perceived accuracy of the information was only sixth, just above how well the branding of the website was recognised. Therefore, these findings show users evaluate credibility along many different dimensions, not just information veracity. The fact the graphical elements are so important is a key point.

The use of multiple dimensions is explained by the fact that trustworthiness and expertise can encompass multiple distinct concepts, as noted by Shah et al. (2015). For instance, some possible components of something deemed as trustworthy are reputation, professionalism, or not involving bias. Expertise can be inferred from something regarded as possessing high quality, accuracy or correctness, or if an authority has approved or created it. These all combine in some manner to cause the viewer to believe the information or information source is credible. Each of these sub-concepts is related to a different dimension or aspect of a website, resulting in many ways to evaluate the extent of a website's credibility. An example would be a heuristic used by users in Fogg et al. (2001) to evaluate trustworthiness. The heuristic involved evaluating website advertising or marketing content. A lack of advertising and marketing implied a lack of bias on part of the website.

Another complication is that the overall credibility can involve several different types. Flanagin and Metzger (2008) propose several categories of credibility for information sources which are summarized in Table 2. Source and information credibility are the main ones. But in a highly networked world, the others should not be ignored when applicable to the type of website or information being assessed.

The last important factor relevant to a full assessment of credibility is characteristics of the viewer themselves. If they are not taken into account, then any attempt to assist in evaluating credibility might be rejected for being unbelievable. Stanford et al. (2002) provide some useful insights. Their study compares how experts in the fields of health and finance assessed website credibility relative to how average consumers did. The results are in line with Fogg et al. (2001) in that consumers primarily used evaluation of superficial website characteristics as heuristics in evaluating credibility. Experts however focused on assessing primarily the information content of the websites. They used their domain specific knowledge to focus on more important things that they, as experts, knew definitely impacted credibility. An example would be the experts' use of their knowledge and assessment of the organization's

Table 2: Type of credibility and Description of the category

Type	Description
Source	Based on the viewer's psychology and subjective assessment of the source
Information	An assessment of more objective measures such as veracity
Media	Related to the type of medium (newspaper, website, TV, radio, etc.)
Conferred	Credibility assigned by those deemed to be experts
Tabulated	Crowd-sourced assessments of credibility, like Facebook likes
Reputed	Reputational based on personal and social networks
Emergent	Credibility emerging through group based work, e.g. Wikipedia

reputation. A consumer might or might not have such knowledge about reputation. As a consequence, the consumer might possibly end up using the quality of the graphic design as some sort of proxy measure for 'professionalism'. Metzger (2007) provides an explanation for this, showing consumers were unlikely to go to the effort of cross-checking information or perform other more thorough procedures for verifying credibility. The overall conclusion is that to effectively assist consumers, they should be provided with information that allows them to readily emulate experts.

3 Categories Of Features

A thorough examination of website characteristics useful in evaluating credibility can be found in the literature review of Shah et al. (2015). Shah et al. use the term *factor* rather than *feature*, but they can be regarded as synonyms. This paper uses feature as a term to stay consistent with section 4, which discusses machine learning (ML) based approaches. Shah's categories are high level ones that are useful from a theoretical or analytical perspective. But they are not useful from a practical perspective, i.e. they denote what features to consider, but not the commonalities in how to process them. The rest of this section discusses features used in previous studies, grouping them into three categories: *extracted*, *proxy*, and *content based*. This categorization will aid in the design of subsystems that extract and analyse features of a specific type.

3.1 Extracted Features

The adjective *extracted* indicates that these features are distinct and encapsulated pieces of information. They are looked for within the web page content – extracted features are not inferred or derived from something else. Two examples would be features based on metadata and checklists. Eysenbach and Diepgen (1998) propose using website metadata, collated and managed by a central authority, to assess credibility. Useful information such as authorship or references to scientific studies would logically be in the metadata. Check-

lists are used in the earliest research into credibility which focused on health and medical related websites. Two notable organizations that have drawn up checklists are HONcode (Boyer et al., 1998) and DISCERN (Charnock et al., 1999). These checklists use specific criteria to evaluate a website. The criteria are straightforward ones such as the presence of authorship credentials, the currency of the information, and the quality of the information as judged by an expert. Checklists designed by experts are a way of establishing conferred credibility. Subsequently, Price and Hersh (1999) propose using software to find, extract, and process specific information relevant to credibility. Two examples of studies using software to extract features are Aphinyanaphongs et al. (2007) and Mavlanova and Benbunan-Fich (2010). Aphinyanaphongs et al. (2007) examine web page content for pseudo-scientific terms (a negative signal for credibility). Mavlanova and Benbunan-Fich (2010) use the presence of contact information or a stated returns policy as a feature for evaluating e-commerce websites.

3.2 Proxy Features

In contrast to extracted features, *proxy features* are ones that are based on technical aspects of a website. Logically, they are not directly associated with credibility. Instead, they can be used to infer some measure of the website's credibility. These features are based on the PageRank, the URLs, and the HTML of the website pages. First, Google's PageRank has frequently been used as a measure of credibility (Griffiths et al. (2005); Aphinyanaphongs et al. (2007); Schwarz and Morris (2011); Olteanu et al. (2013)). Using PageRank is a way of assessing the tabulated credibility of a website. This assumes PageRank is a valid measure of how other websites or Google's systems regard the target website. There have been a number of enhancements to the graph based methodology that PageRank represents; a summary of them can be found in Abbasi et al. (2012).

URL-based features are ones associated with an analysis of the domain name and the URL. Detection of typosquatting (Chen et al., 2009) is the first notable issue. Typosquatting is not usually framed as a credibility related problem, but it does directly relate to the trustworthiness aspect. Another evaluation of trust is verification of HTTPS SSL certificates, which is a function browsers now contain. If the browser detects a mismatch between the domain name and what the SSL certificate says, it will present a page warning to the user that they may be accessing a different website than they intend to. As for URL analysis, Abbasi et al. (2010) is an interesting study into the similarities amongst 'fake' websites. Fake websites are ones that, despite their seeming credibility, defraud their erstwhile customers in some manner. URLs and links in the web pages were processed in simple ways to create some of the features used in the statistical learning process.

Finally, evaluating HTML at a technical level is a possible feature based on the idea it is a measure of the professionalism of the website creator. Wassmer and Eastman (2005) investigate using the number of HTML validation errors reported by the W3C HTML validator (`http://validator.w3.org/`). A more sophisticated way to analyse HTML is found in Abbasi et al. (2010). They discuss how fake websites tend to reuse HTML source code between the pages of copies of fake websites, along with malicious Javascript scripts.

3.3 Content Based Features

The scope of content based features is the widest due to all the ways of processing web page content for useful information. The methods used to create features can be roughly divided into three categories: brute-force, semantics based, and visual features. Brute-force methods do not try to interpret the text. Instead, the text is used as-is, or is processed into n-grams. Semantic approaches, however, create features representing an understanding of the text. Linguistic analysis or text summarization would be examples of this. Finally, the images used by websites are useful features as well as the aesthetic aspects of a website, when quantified.

Olteanu et al. (2013) use the simplest of brute-force features such as the number of nouns in the text or exclamation points. The theoretical rationale for this is unclear. More justifiable measures such as readability and the number of spelling and grammar errors are also used. The presence of advertising, easily determined through HTML analysis, is another simple feature used in many studies. Advertising serves as a negative signal when credible websites are expected to be unbiased or non-profit. Boyer et al. (2015a) compare using stemming and n-grams as features in classifying websites according to the HONcode checklist. Boyer and Dolamic (2016) is a follow-up that evaluates using n-grams for language independent classification.

Wawer et al. (2014) follow up the study of Olteanu et al. (2013) by using a content analysis tool to create vectors of numbers. These vectors represent the psycholinguistic and psychosocial categories of words in the text. Words that invoke trust on part of the reader are differentiated from the ones that invoke distrust. Another semantics based approach is text summarization, used in Balcerzak et al. (2014). However, the results were not promising; humans evaluators correlated informative sentences with credibility, but the algorithm could not reliably extract the most credible sentences.

Images can be used as features in two ways. The first involves no interpretation of them as an image. Abbasi et al. (2010) find website images are useful features because fake websites, being copies of one another, tend to reuse image files. The second way is to assess an image's impact on the viewer. Robins and Holmes (2008) establish that more professionally designed websites are perceived as more credible. Alsudani and Casey (2009) investigate what aesthetic factors have an impact on credibility. Lowry et al. (2014) investigate how logos and site design can influence credibility assessment, when viewers perceive them as relaying one of the concepts behind credibility. Finally, Wu et al. (2011) is an interesting study into how to process a web page into feature vectors. These vectors are based on the HTML layout and graphical elements such as colour and images. With them, a Support Vector Machine classified web pages comparable to human evaluations.

4 Assessment Approaches

Ever since Eysenbach and Diepgen (1998) first proposed using software to assist users in evaluating websites, the research into credibility has only grown. But a review of the litera-

ture shows there are only three categories of approaches from an implementation perspective. These are *manual approaches*, *scoring systems*, and *machine learning based*. This section examines each, discussing how they work and their costs and benefits. These three approaches are complementary to the computer based approaches discussed in Shah et al. (2015).

4.1 Manual Approaches

Manual approaches are ones that involve some component of human effort, i.e. manual labor, to evaluate and collate credible sources of information. Table 3 lists different types of approaches. The benefit is obvious as using human experts ensures the credibility assessment is correct. But as both Eysenbach and Diepgen (1998) and Price and Hersh (1999) explain, there are practical problems to these types of manual approaches. The most significant are the need and cost for experts to evaluate a website and its information. The second involves HCI factors such as the ability or desire of consumers to use a checklist or tool (see Bernstam et al. (2005)). Finally, the third important consideration is the amount of ongoing work to maintain the information, be it metadata, a whitelist of validated websites, or a blacklist of not credible websites.

Despite these limitations, efforts have been ongoing to create databases of validated sources of information. Most of these databases are medically oriented as having credible sources of information is very important for health practitioners and consumers alike. One notable effort is that by the Health On the Net organization (`https://www.healthonnet.org/`). With such medical databases, specialized search engines such as Khresmoi (`http://everyone.khresmoi.eu/hon-search`) allow users to access medical information more credible than the highly questionable information returned through a standard Google search. A more general type of database would be the Web Of Trust database (`https://www.mywot.com/`). The WOT is an interesting manual approach that uses crowd sourcing as a foundation. The browser extension connects the user with a database containing information on how other users have evaluated websites. In effect, the WOT Website Checker utilizes tabulated, reputed, and emergent credibility.

Table 3: Type of effort and Description of effort (Price and Hersh, 1999)

Effort	Description
Self-regulation	Adherence to a set of principles for certifying information (e.g. the HONcode initiative (Boyer et al., 1998))
Checklist	A rating tool for users to use to verify the correctness of information (e.g. DISCERN (Charnock et al., 1999))
Reviews	Third party websites review and rate other sources of information (e.g. MyWOT, https://www.mywot.com/)
Accreditation	Experts review and index credible sources of information (e.g. CliniWeb (Hersh et al., 1996))

4.2 Scoring Systems

A scoring system for assessing credibility is one that evaluates web page features and assigns a score in some fashion. An algorithm then calculates an overall credibility score for the web page. It is a straightforward idea which can be automated, reducing the need for human involvement. But examining the few studies into this type of evaluation raises two fundamental questions. These questions are how to score and weight features properly and how to devise a good overall scoring algorithm.

So what is a logical way of scoring individual features? Price and Hersh (1999) do not describe the exact logic behind how they score features. Instead, they mention that the scoring in their system can be adjusted as required. This implies a need for customization per each type of credibility assessment. The other issue is with the weights of features within the overall scoring algorithm. How would one properly weight different classes of features against each other? Some are binary ones in that they are either present or not while others are continuous variables. An example of these would be author information and a readability score. Which feature is more important and what exactly is a sufficient level of readability depends very much on the type of information and website being evaluated. Much research would be needed to devise a robust framework for weighting features appropriately given a specific context.

The second problem is that every overall credibility scoring algorithm would need customization per application. An example of this is the Automated Quality Assessment (AQA) procedure described in Griffiths et al. (2005). A complex procedure first weights common words and phrases found in an expert validated training set of web pages. Then web pages in a test set are scored using this data and various formulas. The results were satisfactory in that high quality web pages in the test set were categorized correctly. But Griffiths et al. (2005) also review the obvious limitations. These limitations are the human labour costs in creating and evaluating data sets, different health topics would need their own AQA process, and website authors could easily game the AQA score by including high quality terms without actually improving the content. These limitations, when generalized, apply to any overall credibility score algorithm.

4.3 Machine Learning Approaches

Explaining machine learning (ML) in general is beyond the scope of this paper. But essentially, ML approaches to evaluating credibility are similar to scoring systems without their drawbacks. Large amounts of data and the power of statistics help with deciding what features are most important. Many techniques have been invented for handling various types of features in a consistent way. As for the overall scoring, general purpose algorithms justified with mathematical theories replace ad hoc algorithms devised by humans. However, the problem of acquiring good training data for supervised algorithms is still present in terms of the cost and time of using human experts.

The studies reviewed reveal no consensus on what the ideal ML algorithm might be or exactly what features are essential. Support Vector Machines (Aphinyanaphongs et al., 2007;

Abbasi et al., 2010; Sondhi et al., 2012; Wawer et al., 2014) were the most used while Naive Bayes was second (Olteanu et al., 2013; Boyer et al., 2015b; Boyer and Dolamic, 2016). It is not unusual that these two predominate as they are binary classifiers – i.e. a web page is either credible or it isn't. Jaworski et al. (2014) use neural networks and linear regression to assign a credibility score from 1 to 5. Clustering is only used in one study (Liu et al., 2010) which evaluates a system for determining if web pages are phishing pages.

Truth Discovery (Yin et al., 2008; Li et al., 2016) is a subject whose application to evaluating credibility is obvious. It consists of iterating over a set of claims, recalculating confidence levels associated with them until they reach a stable state. A significant advantage is that the algorithm is an unsupervised one. Popat et al. (2016) investigate evaluating the truth of claims with a classifier called Distance Supervision. Their classifier is an extension to truth discovery that handles arbitrary claims that do not need a specific linguistic structure.

5 Discussion

As sections 3 and 4 show, the types of features and approaches used in evaluating credibility can be abstracted to a small set. From an engineering standpoint, this is promising. It implies generic reusable modules can be made instead of customised ones for each situation. Generic modules that take a specific type of input and process it in a focused way are much easier to design, build, and reuse. For example, a module that extracts features will use the same fundamental algorithm no matter the type of information. Processing proxy features would work in the same way. The content processing module would be more complex or consist of a set of submodules. But a good abstraction of the submodule interfaces would result in a flexible plug and play architecture.

The same idea of modules applies to ML approaches. The various open source ML toolkits available would serve as a foundation. The other two approaches (manual and scoring systems) are not relevant in terms of automating credibility evaluations. However, the validated databases created from other projects like Health on the Net would certainly be – a module that checks these databases for information on the website being evaluated is an easily implemented idea. Subsection 6.2.3 contains details of these generic modules.

However, the literature review has revealed the type of website being evaluated is a significant factor in how to evaluate its credibility. This is because the set of features that need to be examined is different for each type of website. Table 4 lists feature categories from (Shah et al., 2015). It is a useful guide for deciding how to analyse a website and what features to consider. Because multiple distinct concepts comprise credibility, as explained in section 2, different types of websites should be evaluated with a different emphasis placed on each category. Table 5 provides some examples of types of websites, the feature categories that are most important for that type, and specific examples of relevant features. The conclusion from this is two-fold. First, a website being evaluated must be categorised (e.g. health/medical, financial, e-commerce, news). Second, the general purpose modules would need to be dynamically configurable so that they examine the raw data and process it as required depending on the type of website.

Table 4: Categories of features and Explanations of them (Shah et al., 2015)

Category	Explanation
Accuracy	The correctness of the information presented on the website
Authority	The perceived or provable experience, reputation, or certification of the author or source of information
Aesthetics	The extent to which the combination of colours, layout, images, videos, fonts, etc. used result in a well designed web site
Professionalism	A perception of the organization and how well the website is managed, maintained, or of its policies
Popularity	The website, organization, or author's reputation along with measures such as PageRank
Currency	The frequency of updates applied to the content or the timeliness of it
Impartiality	The lack of bias in the content, e.g. a determination of the website's intent
Quality	An assessment of factors that depend on the exact type of content (e.g. no spelling mistakes in a news website article)

The type of website is also related to the goal or purpose of establishing credibility. As mentioned in subsection 3.2, the trustworthiness aspect of credibility applies to security concerns and websites. The research into detecting domain typosquatting, fraudulent websites, and phishing rarely references the concept of credibility because they don't evaluate

Table 5: Type of Website, Feature Categories, Example Features

Type of Website	Feature Category	Example Features
Health/Medical	Authority	Author and organization accreditation
	Accuracy	Checklist score
	Quality	Expert assessment score
E-Commerce	Professionalism	Privacy and return policies stated
	Aesthetics	Graphic design quality score
Financial	Authority	Brand recognition and reputation
	Professionalism	Clear explanation of services and limitations
	Impartiality	No explicit sales pitches
Blog	Popularity	Author reputation and blogs that link to it
	Quality	Quality of writing (e.g. readability score)
	Authority	Perceived level of knowledge about subject

the expertise aspect of credibility. But any system that does can apply the same general functionality towards the goal of detecting these problems. In this regard, credible becomes a synonym for trustworthy. How this affects the user interface for a system that assists users is discussed in subsection 6.1.

6 Proposed System Design

The idea of a system that assists users in evaluating website credibility is not a new one. Wang and Liu (2007) and Schwarz and Morris (2011) investigate how to assist users, focusing specifically on UI issues and on an end-to-end system for a specific type of website. This paper, inspired by those two studies, proposes a system called DIOGENES. It is designed to handle any type of website while also assisting users in evaluating websites. Section 6.1 describes the user interface, implemented as a browser extension. Section 6.2 is an overview of the back end which both acts as a central repository of information and performs the evaluation process. The ideal system would consist of a browser with an inbuilt capability of evaluating websites. However, there are practical reasons for a client-server architecture, which are performance and collation of evaluation data.

6.1 User Interface

There are three choices for how to implement DIOGENES from the user interface perspective. The first is a new browser with extra functionality; subsection 6.2 discusses this idea's drawbacks. The other two are a search portal and a browser extension which are examined in the rest of this section. Both have a greater probability of user acceptance than a new brand of browser, but a browser extension affords more utility because of the anticipated use case. This use case consists of the user starting up their browser and browsing from web page to web page. In the background, the extension fetches data on the web pages' credibility from DIOGENES. When the user wants more detail on a web page's credibility, they can invoke a new function that displays a page providing them with more information.

A search portal of the kind described in Wang and Liu (2007) is a straightforward and practical tool. It consists of a Java based application that accepts user queries on health and medical related topics. The output is a list of web pages related to the query and an evaluation of their credibility based on specific criteria. The downside is that it is a separate application – another tool the average user would need to use besides a browser. A search portal also precludes web pages reached through casual or random browsing from evaluation; the user has to be explicitly searching for information. Therefore, integrating the functionality into a browser makes more sense from an HCI perspective; web page evaluations can be an ongoing process in parallel with the user's primary focus of surfing the Internet.

The WebOfTrust browser extension (`https://www.mywot.com/`) and the one developed by Schwarz and Morris (2011) are excellent models to base DIOGENES on. The user would only need to install an extension for their browser, so there are no drawbacks such as using

a new browser or a different tool. Figures 1, 2, and 3 show the different modes the browser could be in, depending on the website and the user's actions.

Figure 1 shows the normal mode of the browser. In this mode, the user is browsing the Web, going from website to website. There is an icon in the top right controlled by DIOGENES' browser extension. In figure 1, this icon is represented by a question mark. Another possible UI option would be an overlay at the bottom of the screen, much like OS X's Notifications pullout or the Dock. User experience testing would establish what is the best design. As the user traverses web pages, the state of this icon would change. This new state would be a summary of credibility related information for the current web page, e.g. the question mark in black means there is credibility information available. A grey icon would mean no credibility assessment is currently available. By clicking on the icon, the user would then be presented with a page that displays details about the evaluation of the web page's credibility (see figure 2). The important thing to note is that a colour coded system for indicating credibility (e.g. Red, Yellow, Green) is not viable. The rationale for this is explained below. Instead, the state of the icon would shift to indicate there was credibility information available or not. Thus the user is not distracted from the main task of browsing the Web.

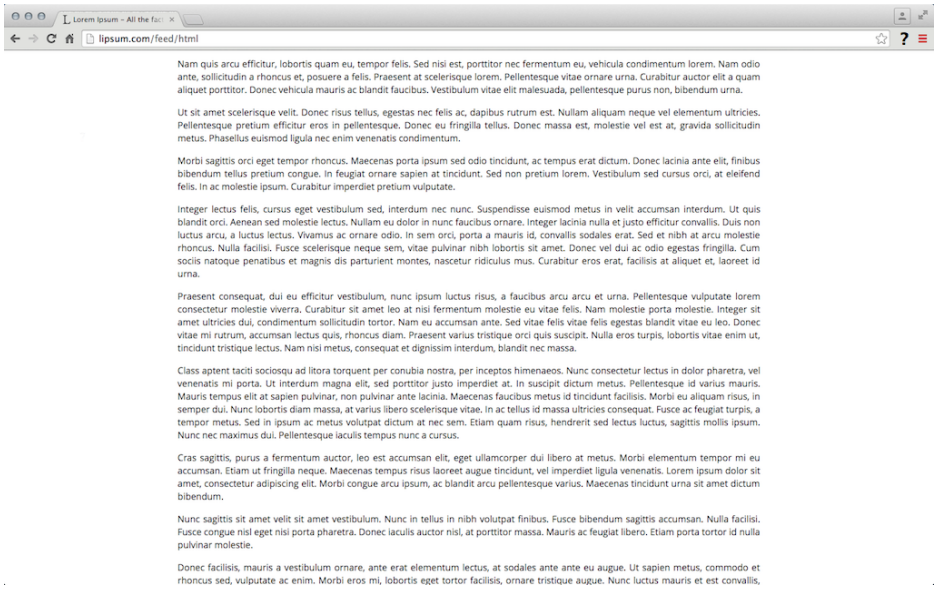

Figure 1: Example of normal web browsing mode

Figure 2 is an example of the page shown when the user requests information on the current web page's credibility. The back end of DIOGENES would provide the information to the browser extension. Exactly what information is displayed would depend on the type of website as each has a different set of important features relevant to credibility. Exactly what these features are and how to present them requires further research. The overall goal, however, is the same for each type of website. It is not to provide the user with an absolute determination of credibility, along the lines of a scale going from Red to Green or a number

from 0 to 100 percent. Instead, the goal is to provide them with the information they need to evaluate the web page in the same manner as an expert. This has a two-fold purpose. One, it would encourage users to evaluate the website in a more comprehensive way and not to be distracted by surface level assessments like graphic design. Two, it would prevent accusations of bias from being levelled against the operators of DIOGENES. Information is presented to the user, but not a conclusion about that information. Thus they are free to evaluate overall credibility as they see fit based on their own standards. This is very important when it involves subjects where bias is an notable factor, political news being the most prominent of these.

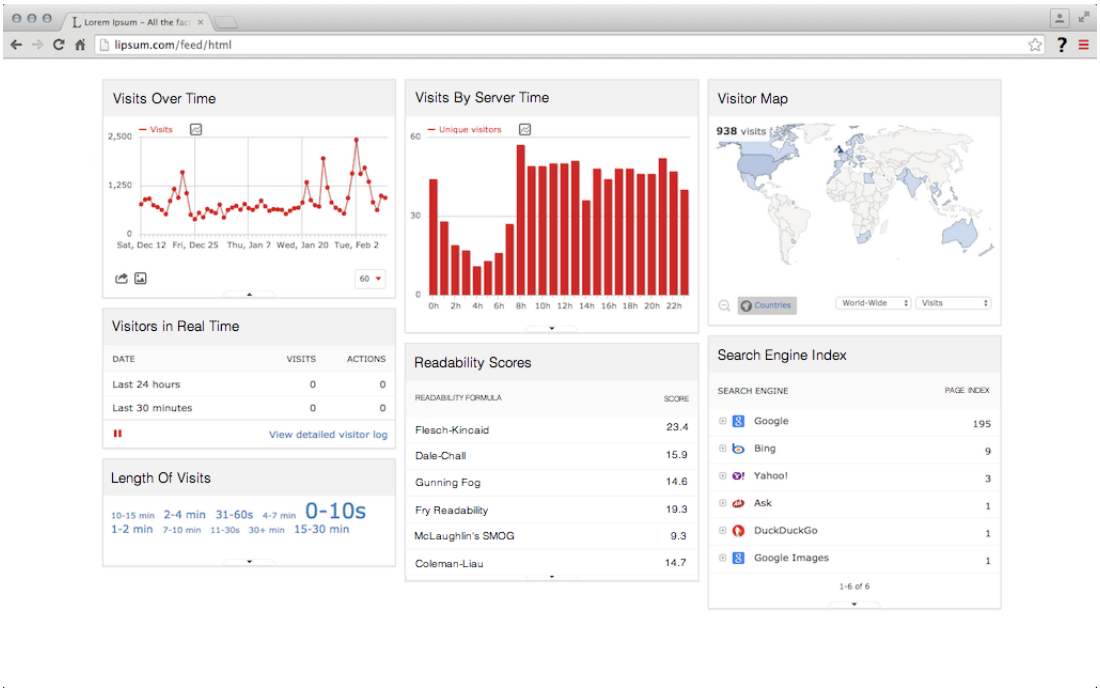

Figure 2: Example of credibility assessment measures

Figure 3 is an example of the page shown when the user navigates to a URL that DIOGENES determines is not trustworthy. This screen is analogous to the ones browsers show when they determine there is a basic security issue. In effect, browsers already have a basic credibility evaluation system built in. It evaluates SSL certificates for websites that use the HTTPS protocol. If there is an issue with the SSL certificate, the browser displays a warning screen to the user explaining the situation. More advanced users have the option to continue to the website if the browser allows this. Additionally, the Chrome browser also has support for blocking the user from sites Google has determined are serving malware to users. This functionality is exactly what DIOGENES would offer for more advanced types of threats - phishing sites, fraudulent websites, etc.

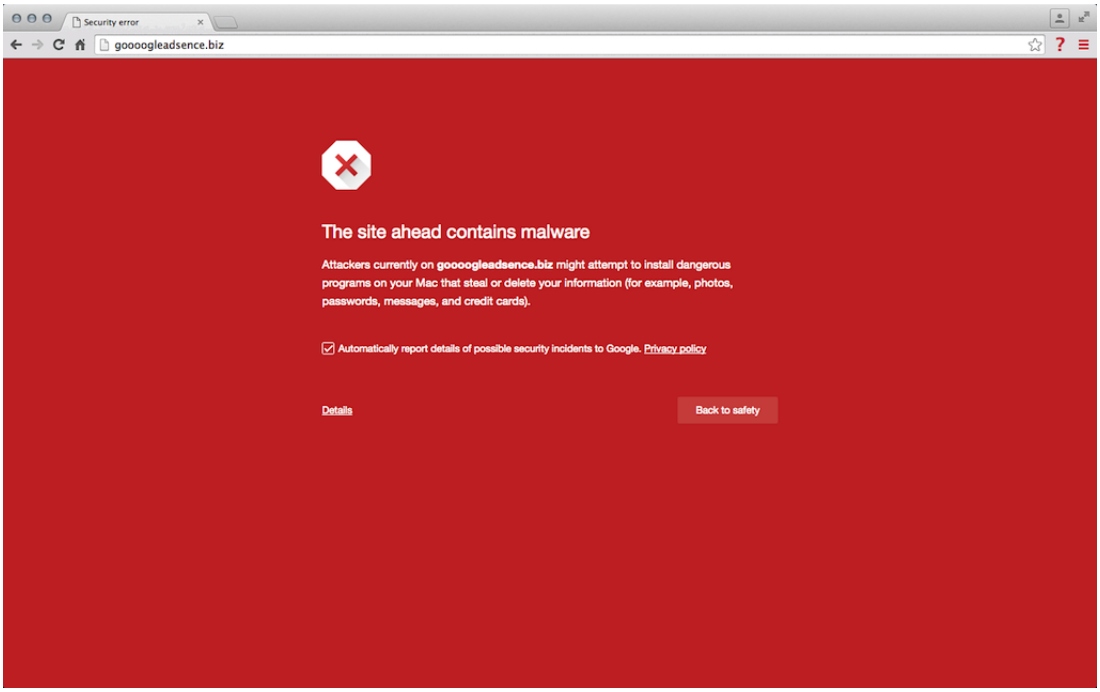

Figure 3: Example of a security alert screen

6.2 Back End

If a browser could evaluate websites and pages in real time, users could get credibility assessments immediately. But as always, there are limitations and tradeoffs. A consumer's typical home computer does not have sufficient capabilities to perform real time evaluations. Additionally, a large portion of the software's functionality would need to be built from basic components, especially the machine learning modules. The application would be large and complex. The last consideration is user acceptance – would users want to install and use a new browser over the one they already prefer?

The benefits of a client-server type architecture are more compelling. A centralised server based on existing products such as Amazon's Machine Learning (`https://aws.amazon.com/machine-learning/`) or the web scraping service Scrapy (`https://scrapy.org/`) is far more efficient. This approach also reduces or simplifies the work required because it becomes predominantly integration rather than custom development. Another benefit of a centralised system is that evaluations performed for one user can be stored for when another user requests information on that website. This is the same strategy used by the WebOfTrust system (`https://www.mywot.com/`).

Figure 4 shows the back end architecture of DIOGENES. The arrows represent the data flows of processing an example URL; not every possible module or path is shown. The fundamental process is a pipeline. The first step is pre-processing, then feature assessment in parallel. Some feature assessments may take more time than others, so the next stage is a buffering process that stores incomplete results. When all the relevant features for the page

have been assessed, they are provided to a machine learning module that classifies the web page as desired. The web page credibility database stores the feature and credibility assessments. Each of these major subsystems is described in more detail below. Web pages that create errors in any of the stages are logged for human review and a result of 'indeterminate' would be added to the credibility database. When the cause of the error has been diagnosed and fixed, this temporary record for the web page would be removed.

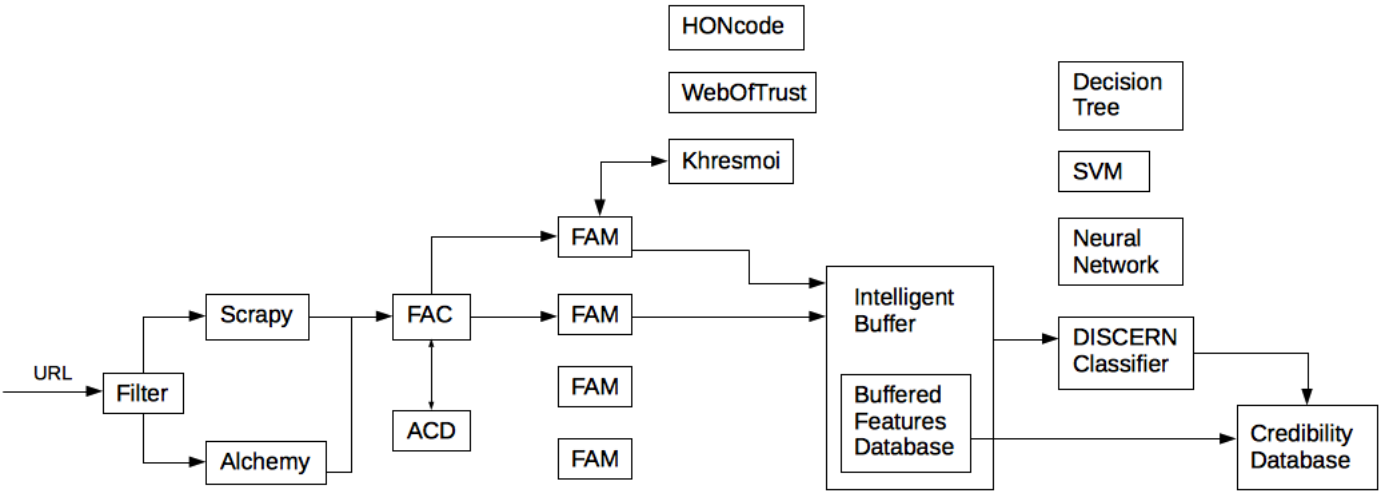

Figure 4: DIOGENES back end architecture

6.2.1 Credibility Database

The credibility database is a NoSQL database that holds web page URLs and associated credibility related information. The type of information stored will vary from URL to URL depending on the type of website and the set of features assessed. SQL operations such as relational joins are not needed; the only operation that is, are lookups on the URL. In future, more complex lookups on fields other than the URL might be useful for certain types of research.

6.2.2 Preprocessing Modules

The pre-processing step consists of filtration, fetching the web page, and determining the type of website or web page. First, when a request to evaluate a URL is received, a simple filter would validate it for errors or other problems. The filter would then look up the URL in the credibility database to see if it has already been assessed. If the URL has been, then the credibility information would be sent back to the client immediately. If it hasn't, then a result of 'unknown' would be sent back to the client and the URL sent to the other two steps of pre-processing. Due to the nature of the Web, clients cannot be put on hold to wait for the pipeline to fully assess a web page. Subsequent clients would receive the stored results

of the URL's credibility assessment. To alleviate this problem as much as possible, a web spider could be used to search for new URLs to process, thus pre-populating the credibility database.

The next two steps are done in parallel by one module, using the Scrapy (`https://scrapy.org/`) and Alchemy (`http://www.alchemyapi.com/`) services. Scrapy is a framework for web scraping; this service would fetch the web page content and pre-process it as needed. Alchemy is a set of tools for semantic text analysis; it would serve as the foundation for determining the type of website and type of information being assessed. The combined output of these services is then sent to the feature assessment controller described in the next section. As there are no data dependencies in these steps, horizontal scaling by increasing the number of pre-processing modules would be possible.

6.2.3 Feature Assessment Modules

There are three main components to this step of the pipeline. They are the feature assessment controller (FAC), the assessment configuration database (ACD), and the individual feature assessment microservices (FAMs). The input to the FAC is the web page contents as scraped by Scrapy and the semantic data from the Alchemy service. The semantic data would allow the FAC to determine the type of website or the subject of the web page being assessed. This topic information would be used as an index into the ACD. The ACD would correlate topics to the features to consider and the parameters used by the FAMs. With the configuration data from the ACD, the FAC would fire off a set of tasks to the corresponding FAMs. These tasks are implemented through a distributed task queue built using a library such as Celery (`http://www.celeryproject.org/`). As each task finishes, it would report its results to the Intelligent Buffer service described in the next section.

Most of the FAMs would have straightforward functions, such as examine the web page content for a specific kind of information (e.g. accreditation information for medical web pages). The microservices architecture would result in a collection of simple and easily developed web services. But there is one type of FAM that should be specifically mentioned. It is a service that checks external sources of credibility information such as the WebOfTrust database, the Health On the Net database, or the Khresmoi search engine.

6.2.4 Intelligent Buffer

The foundation for the Intelligent Buffer would be a NoSQL database. NoSQL is a requirement because a variety of schemas are needed to store the feature related information. The key fields common to any record are Job ID, Update Time, and a list of Sub-job IDs. When the FAC starts a new job, it would create a record in this database. The Update Time would be the time the job started. The sub-job IDs are the IDs of the FAM related tasks. As each task finishes, it would notify the Intelligent Buffer service. The Buffer would then look up the record in the database, add the result of the FAM task, update the update time, and mark the sub-job as finished. If all the sub-jobs have finished, the Buffer would pass the feature information to the machine learning stage. Horizontal scaling would be possible by

increasing the number of buffer services running. If simultaneous updates are not handled efficiently, then a job queue would allow operations to be serialized. Error handling would consist of running a background process that examines update times. If the current time minus a record's update time exceeds a limit, then the job would be labeled as failed and the error logged for evaluation and debugging.

6.2.5 Machine Learning Modules

The foundation for these modules would be the numerous libraries and ML systems already available, e.g. the scikit-learn library (`http://scikit-learn.org/`) and Amazon's Machine Learning service (`https://aws.amazon.com/machine-learning/`). Other sources would be the classifiers developed by Allam et al. (2016) and Boyer et al. (2015a). As the modules are well encapsulated ones, horizontal scaling would be possible using Amazon Web Services or the equivalent. Most of the work involved would consist of developing interfaces to the existing tools. The input interface would translate the incoming feature data to the format required by the tool. The output interface would transform the machine learning results into the format required by the credibility database. Further investigation would determine if custom interfaces would be needed for each tool or if a generic set can be developed. The interfaces would also perform error handling as needed when the particular tool reports an input or output error.

7 Conclusion

Assessing website credibility is a complex topic that touches upon many subjects such as information science, communication science, psychology, computer science, and human-computer interaction. The proposed system, DIOGENES, is one based on an effort to synthesise all the different research studies. The main aspects considered are the features to use, the approaches to assessing credibility, the type of website being evaluated, and the goal of establishing credibility. By finding the commonalities from an engineering perspective, a set of well defined and encapsulated modules was designed that can serve as building blocks for an automated system. The commercial utility of such a system is questionable, as the business model is uncertain. But DIOGENES, if open sourced, could certainly serve as a reusable testbed. This testbed would be useful for more extensive research into how users evaluate the credibility of a website, especially considering the growing problem of not credible news sources and stories.

References

Abbasi, A., Zahedi, F. M., and Kaza, S. (2012). Detecting Fake Medical Web Sites Using Recursive Trust Labeling. *ACM Trans. Inf. Syst.*, 30(4):22:1–22:36.

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., and Nunamaker, J. (2010). Detecting Fake Websites: The Contribution of Statistical Learning Theory. *MIS Quarterly*, 34(3):435–461.

Allam, A., Schulz, P. J., and Krauthammer, M. (2016). Toward automated assessment of health Web page quality using the DISCERN instrument. *Journal of the American Medical Informatics Association*.

Alsudani, F. and Casey, M. (2009). The Effect of Aesthetics on Web Credibility. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, pages 512–519. British Computer Society.

Aphinyanaphongs, Y., Aliferis, C., et al. (2007). Text categorization models for identifying unproven cancer treatments on the web. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 968. IOS Press.

Balcerzak, B., Jaworski, W., and Wierzbicki, A. (2014). Application of TextRank algorithm for credibility assessment. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01*, pages 451–454. IEEE Computer Society.

Bernstam, E. V., Shelton, D. M., Walji, M., and Meric-Bernstam, F. (2005). Instruments to assess the quality of health information on the World Wide Web: what can our patients actually use? *International Journal of Medical Informatics*, 74(1):13–19.

Boyer, C. and Dolamic, L. (2016). N-gram as an alternative to stemming in the automated HONcode detection for English and French.

Boyer, C., Dolamic, L., and Falquet, G. (2015a). Language Independent Tokenization vs. Stemming in Automated Detection of Health Websites? HONcode Conformity: An Evaluation. *Procedia Computer Science*, 64:224 – 231.

Boyer, C., Dolamic, L., Ranasinghe, M., and Baujard, V. (2015b). An automated HONcode detection system informs internet users of HONcode compliance. *Swiss Medical Informatics*, 31.

Boyer, C., Selby, M., Scherrer, J.-R., and Appel, R. (1998). The Health On the Net Code of Conduct for medical and health Websites. *Computers in Biology and Medicine*, 28(5):603–610.

Charnock, D., Shepperd, S., Needham, G., and Gann, R. (1999). DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology and Community Health*, 53(2):105–111.

Chen, G., Johnson, M. F., Marupally, P. R., Singireddy, N. K., Yin, X., and Paruchuri, V. (2009). Combating typo-squatting for safer browsing. In *Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on*, pages 31–36. IEEE.

Credibility (n.d.). *The American Heritage Dictionary of the English Language, Fourth Edition*. Retrieved from http://www.dictionary.com/browse/credibility.

Eysenbach, G. and Diepgen, T. L. (1998). Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. *BMJ*, 317(7171):1496–1500.

Flanagin, A. J. and Metzger, M. J. (2008). Digital Media and Youth: Unparalleled Opportunity and Unprecedented Responsibility. *Digital Media, Youth, and Credibility*, pages 5–27.

Fogg, B., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., et al. (2001). What Makes Web Sites Credible?: A Report on a Large Quantitative Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 61–68. ACM.

Fogg, B. and Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 80–87. ACM.

Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. (2003). How Do Users Evaluate the Credibility of Web Sites?: A Study with over 2,500 Participants. In *Proceedings of the 2003 Conference on Designing for User Experiences*, DUX '03, pages 1–15, New York, NY, USA. ACM.

Griffiths, K. M., Tang, T. T., Hawking, D., and Christensen, H. (2005). Automated assessment of the quality of depression websites. *Journal of Medical Internet Research*, 7(5):e59.

Hersh, W. R., Brown, K. E., Donohoe, L. C., Campbell, E. M., and Horacek, A. E. (1996). CliniWeb: managing clinical information on the World Wide Web. *Journal of the American Medical Informatics Association*, 3(4):273–280.

Jaworski, W., Rejmund, E., and Wierzbicki, A. (2014). Credibility Microscope: Relating Web Page Credibility Evaluations to Their Textual Content. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 297–302.

Kakol, M. and Nielek, R. (2015). What affects web credibility perception? An analysis of textual justifications. *Computer Science*, 16(3):295–310.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2016). A Survey on Truth Discovery. *SIGKDD Explor. Newsl.*, 17(2):1–16.

Liu, G., Qiu, B., and Wenyin, L. (2010). Automatic detection of phishing target from phishing webpage. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 4153–4156. IEEE.

Lowry, P. B., Wilson, D. W., and Haig, W. L. (2014). A picture is worth a thousand words: Source credibility theory applied to logo and website design for heightened credibility and consumer trust. *International Journal of Human-Computer Interaction*, 30(1):63–93.

Mavlanova, T. and Benbunan-Fich, R. (2010). What Does Your Online Pharmacy Signal? A Comparative Analysis of Website Trust Features. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10.

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091.

Netcraft Survey (2016). Number of websites. Accessed: 2016-11-01. (Archived by WebCite at http://www.webcitation.org/6lhJlHtez).

Olteanu, A., Peshterliev, S., Liu, X., and Aberer, K. (2013). Web credibility: Features exploration and credibility prediction. In *European Conference on Information Retrieval*, pages 557–568. Springer.

Popat, K., Mukherjee, S., Strötgen, J., and Weikum, G. (2016). Credibility Assessment of Textual Claims on the Web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16. ACM.

Price, S. L. and Hersh, W. R. (1999). Filtering web pages for quality indicators: an empirical approach to finding high quality consumer health information on the world wide web. In *Proceedings of the AMIA Symposium*, page 911. American Medical Informatics Association.

Robins, D. and Holmes, J. (2008). Aesthetics and credibility in web site design. *Information Processing & Management*, 44(1):386 – 399. Evaluation of Interactive Information Retrieval Systems.

Schwarz, J. and Morris, M. (2011). Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1245–1254. ACM.

Shah, A. A., Ravana, S. D., Hamid, S., and Ismail, M. A. (2015). Web credibility assessment: affecting factors and assessment techniques. *Information Research*, 20(1):20–1.

Sondhi, P., Vydiswaran, V. G. V., and Zhai, C. (2012). *Reliability Prediction of Webpages in the Medical Domain*, pages 219–231. Springer Berlin Heidelberg, Berlin, Heidelberg.

Stanford, J., Tauber, E. R., Fogg, B., and Marable, L. (2002). Experts vs. online consumers: A comparative credibility study of health and finance Web sites. *Consumer Web Watch Research Report*.

Steiner, P. (1993). On the Internet, nobody knows you're a dog. *The New Yorker*. Accessed: 2016-11-12. (Archived by WebCite at http://www.webcitation.org/6lyCqdAGu).

Wang, Y. and Liu, Z. (2007). Automatic detecting indicators for quality of health information on the Web. *International Journal of Medical Informatics*, 76(8):575–582.

Wassmer, M. and Eastman, C. M. (2005). Automatic evaluation of credibility on the Web. *Proceedings of the American Society for Information Science and Technology*, 42(1).

Wawer, A., Nielek, R., and Wierzbicki, A. (2014). Predicting Webpage Credibility Using Linguistic Features. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 1135–1140, New York, NY, USA. ACM.

Wu, O., Chen, Y., Li, B., and Hu, W. (2011). Evaluating the Visual Quality of Web Pages Using a Computational Aesthetic Approach. In *Proceedings of the fourth ACM international conference on Web Search and Data Mining*, pages 337–346. ACM.

Yin, X., Han, J., and Yu, P. S. (2008). Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808.