Towards a Knowledge-Aware AI
A. Dimou et al. (Eds.)

doi:10.3233/SSW220005

Evaluating Web Content Using the W3C Credibility Signals

León Viktor AVILÉS PODGURSKI [a,1], Karolina ZACZYNSKA [b] and Georg REHM [b]
[a] *Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany*
[b] *DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany*

Abstract. The credibility and trustworthiness of online content has become a major societal issue as human communication and information exchange continues to evolve digitally. The prevalence of misinformation, circulated by fraudsters, trolls, political activists and state-sponsored actors, has motivated a heightened interest in automated content evaluation and curation tools. We present an automated credibility evaluation system to aid users in credibility assessments of web pages, focusing on the automated analysis of 23 mostly language- and content-related credibility signals of web content. We find that emotional characteristics, various morphological and syntactical properties of the language, and exclamation mark and all caps usage are particularly indicative of credibility. Less credible web pages have more emotional, shorter and less complex texts, and put a greater emphasis on the headline, which is longer, contains more all caps and is frequently clickbait. Our system achieves a 63% accuracy in fake news classification, and a 28% accuracy in predicting the credibility rating of web pages on a five-point Likert scale.

Keywords. content credibility, credibility assessment, credibility signals, content curation, fake news, NLP

1. Introduction

The digital age continues to transform the ways humankind lives, interacts and communicates. As more and more online information is published and consumed, the trustworthiness of web content has become a major societal issue. Fraudsters, trolls, political activists and state-sponsored actors disseminate misinformation and other unreliable and malicious content commonly described as "fake news" [1]. The proliferation of low-quality information on the web is facilitated by its decentralised and nonrestrictive nature, which enables the publication of content without the limitations or qualitative controls associated with traditional media. In a global study, 62% of respondents felt that fake news was prevalent on online websites and platforms [2].

This work focuses on automatically predicting the credibility of web pages' content. Following the definitions of Navok et al. and Grupta et al. we define credibility as the property of being trusted, i. e., a text is declared credible if a user, e. g., based on common sense, believes that the information it contains is to some extent credible [3,4]. Amidst growing concerns over unreliable online information and fake news, a number

[1] E-mail: lv.avilesp@yahoo.de, karolina.zaczynska@dfki.de, georg.rehm@dfki.de

of stakeholders have called for the development of tools, frameworks and technologies to support human credibility judgements on the web [5,6,7,8]. Automated tools can audit information almost instantly and may be deployed for each individual user to flag their consumed content with credibility indicators in real-time. Furthermore, it is vital that users are able to retrace why a given text is classified as (not) credible, and not just given by a black box model trained on a large data set. Previous research has highlighted discriminating structural and linguistic characteristics of web pages with low credibility, such as an increased emotionality [9], shorter average words [10], and less coherent texts [11]. Such properties can be framed as *credibility signals*, measurable units of information that may be used as credibility-indicating heuristics in the credibility assessment process. Each signal represents a certain trait of a web page and may contribute to or detract from its overall credibility. The World Wide Web Consortium (W3C) Credible Web Community Group [12] has published an extensive list with the specifications of more than 200 credibility signals[2]. Studies have shown repeatedly that superficial web page features, like appearance or usability, have a considerable impact on credibility [13,14]. However, we decided to build a credibility assessment system focusing on signals related to the language and actual content of web pages. If our credibility scores are based on properties intrinsic to the evaluated information, the scores will be consistent across platforms for the same content, and should also be much more resistant to adversarial attacks as changes to the content are more costly than modifications of surface characteristics.

The main contributions of this paper are as follows: we present the design and implementation of a software system which calculates the credibility score of a web page through the analysis of (mostly) language- and content-related credibility signals. Although it is a rather popular approach in automated credibility assessment, we decided not to train a machine learning (ML) model for the task. To the best of our knowledge, we are the first to systematically evaluate a subset of the W3C web credibility signals; we develop a web page parsing module and separate evaluation pipelines for a total of 23 credibility signals. Evaluating which credibility signals are particularly relevant to automated credibility assessments, we see that emotional characteristics, various linguistic, especially morphological and syntactical properties, and exclamation mark and all caps usage are particularly indicative. Less credible pages have more emotional, shorter and less complex texts, and put a greater emphasis on the headline, which is longer, contains more all caps and is frequently clickbait. Although there is still ample room for iterative improvements of our credibility evaluation system, we show that a content-focused automated credibility evaluation that is transparent, more robust, domain-independent, modular, and easily expandable is feasible and states a clear added value to existing black-box ML or qualitative approaches. Our system can aid internet users and publishers in their credibility assessments to make more informed credibility decisions about the information at hand; credibility researchers can use the tool to analyse and learn more about web credibility and elements that affect it, as well as study the credibility of specific websites or platforms. The paper is structured as follows: First, we present related work and related sub-fields like fake news (Section 2). Section 3 presents the selected signals and tools used in our system (Section 3.1) and explains how we combine and weight the signals to compute the credibility score (Section 3.2). Section 4 presents the datasets we conducted our test runs on and Section 5 discusses the influence of individual credibility signals for the overall score and the performance of the system.

[2] https://credweb.org/signals-20191126. URLs were all last accessed on 2022-01-21.

2. Related Work

Systems that analyse information, helping users regarding different points of view and previously overlooked indicators of credibility, can lead to more holistic and informed credibility judgements. Automated tools can audit information much faster and at larger scale than human reviewers or fact-checkers, while taking features into account that are difficult for humans to evaluate [5,15]. Chen et al. [5] and Lazer et al. [7] suggest a hybrid approach for addressing the problem of fake news: promoting public literacy and critical thinking for the digital space with initiatives and training, such that individuals may consume online content more consciously and validate and cross-check information themselves; and through automated evaluation and verification systems, employed by users or platforms to support credibility assessments and flag suspicious content. To help users recognise filter bubbles, false news or abusive content more easily, Rehm [8] proposes a decentralised infrastructure on top of the World Wide Web, including corresponding metadata standards, smart content and semantic content enrichment following the main principles of the Semantic Web. Our credibility assessment tool can be perceived as one building block of such an infrastructure, aggregating credibility values that can be made available, for example, as Linked Data or as Web Annotations, for example. Horne et al. [16] introduce an open source toolkit intended to facilitate the systematic exploration of the online news ecosystem, where users may consume news articles and simultaneously receive indicators of their reliability. The W3C Credible Web Community Group has published an analysis of the factors, stakeholders and possibilities for improving web credibility assessments, describing promising technical approaches for each of the credibility assessment strategies *inspection*, *corroboration*, *reputation* and *transparency* [6]. Furthermore, they published a list of more than 200 credibility signals intended to support the creation of interoperable credibility tools by researchers and software developers [12], which we utilise as the foundation for the design of our system.

Most modern credibility evaluation systems use machine learning and Natural Language Processing (NLP) to analyse the credibility of web pages. Olteanu et al. [17] apply statistical tests to determine 22 particularly important web page features as indicators for credibility (from 37 features identified in the literature), to build an automated credibility evaluation framework. Wawer et al. [18] aim to improve the system by Olteanu et al. [17] through leveraging psychosocial and psycholinguistic cues, and analysing word occurrences using bag-of-words models. Investigating words connected to certain trust levels, they conclude that consumers assign lower credibility to financial services and content generated by users or related to borrowing money, while government- and safety-related content was associated with words implying high trust. Esteves et al. [19] compute credibility ratings after analysing lexical and textual properties, as well as groups of HTML tag occurrences and extracting source reputation cues. Giachanou et al. [9] concentrate on emotional signals, proposing an LSTM model that evaluates textual and emotional indicators for credibility assessment, and determine that the inclusion of emotional signals can improve the performance of credibility assessment systems.

Several publications in automated credibility assessment and fake news detection focus on content-related or linguistic properties of online information. Horne and Adalı [10] evaluate psychological, complexity and stylistic features of legitimate and fake news articles and find significant differences in the language of title and text. They conclude that fake news articles attempt to include all central information in the title, while the

text body generally contains only marginally more information. Afroz et al. [20] show that a deceptive writing style can be recognised using stylometry; while authors can intentionally alter their writing style to avoid identification through such an analysis, the actual obfuscation can be detected. Rashkin et al. [21] find that fake news contain more words related to exaggeration, and real news more words related to concrete figures. O'Brien et al. [22] employ a deep neural network for fake news detection based on language features, and discover patterns of language bias, exaggeration and strong rhetoric as corresponding to fake news. Przybyła [23] evaluates a classifier based on stylometric features, and determines sensational and affective vocabulary to be a discriminating factor. Compared to such machine-learning-based credibility evaluation approaches producing one overall score or a binary assessment, our paper follows a different line of argumentation. We calculate the credibility of a web page and evaluate more than 20 credibility signals, defined as small, measurable units of information that may be utilised as credibility-indicating heuristics in the credibility assessment process.

Fake news detection is closely related to, and arguably a sub-domain of automated credibility evaluation, representing a binary classification of the reliability or trustworthiness of web content instead of rating its credibility on a scale. Similarly to automated credibility assessment, fake news detection systems typically utilise machine learning approaches including neural networks to avoid the usage of handcrafted features [1].

3. Methodology

3.1. Signal Selection

In order to build a system which processes web pages and assigns credibility scores, we must identify, extract and analyse relevant information from these web documents. We analysed the list of more than 200 credibility signals by [12] to decide which characteristics of web pages are important for a credibility evaluation and, thus, to be included in our system. Human credibility assessments of web content are often linked to surface signals. Indeed, website appearance, usability and design have frequently been shown to correlate with credibility [24]. We focus on the content and language of web pages rather than surface characteristics for multiple reasons. Several scientific works have focused on language features for credibility classification and fake news detection [10,22,23]. Similarly, we want to investigate the relationship between the credibility of text-based web content and its linguistic features, like a text's vocabulary, tonality, grammar, style and structure. Additionally, we prefer to rely on more robust signals connected to the actual content, rather than superficial indicators. Our credibility ratings are linked strongly to the evaluated data, with ratings being consistent across platforms if the same content is evaluated, i. e., our scores will not change depending on the website that the content was published on. Moreover, precise changes in individual content pieces are more expensive than adapting a website's appearance; as we expect our content-related signals to correlate with credibility and therefore with quality, content providers would have to actually increase the quality of their content to improve its credibility score, and could not just manipulate it by modifying surface properties. Nevertheless, we do include some signals not related to language but pertaining to other characteristics. In the context of the assessment strategies described by the W3C Credible Web Community Group, the

credibility evaluation performed by our system is classified as *inspection* [12]. As it is mainly designed to analyse language, it performs best on text-centric content such as news articles and blog posts.

The signals from the W3C Credible Web list were selected according to compliance with the factors *relevance* (expected correlation between signal values and the credibility of the content, based on previous scientific work), *measurement difficulty* (expected difficulty in automatically extracting the needed information from a web page and determining the signal value), *manipulability/feedback risk* (signals should be sufficiently hard to attack, i. e., it should be disproportionately expensive to modify evaluated content to improve the signal evaluation), and *interoperability* (signal evaluation should reach the same result even when performed by several independent systems) [12]. The metrics we opted to use and evaluate are:

Author: This signal represents whether a web page explicitly states the author of its content through a byline or other means. It has been shown that the presentation of author credentials is positively linked to credibility [14,25,26,27]. We implement the signal's automated evaluation by either extracting the author's credentials directly from the web page's HTML code or through the use of a third-party parsing library.

Clickbait headlines are deliberately sensationalist and misleading in an attempt at enticing consumers to click through to the linked article or content. Clickbait has been found to be indicative of web content with poor quality and low credibility, and is frequently linked to fake news [28,29,1]. Karadzhov et al. [30] construct a fake news and clickbait classifier using stylometric, lexical, grammatical and semantic features with a neural model, which we used for our system.

Grammar & Spelling errors are perhaps the most basic language-related form of amateurism, and intuitively indicate lower quality content. Previous studies have found a correlation between such errors and decreased credibility [31]. To determine the amount of spelling and language errors in the headline and text, we utilise the LanguageTool library. A significant problem here was the large amount of false positives, and we perform extensive filtering of the error matches to exclude errors such as using allegedly wrong forms of quotes or dashes, lexical redundancy, and British English orthography. We also conduct entity recognition using spaCy to avoid classifying names as errors, as these were the majority of error matches.

Language structure encompasses several signals related to morphological and syntactical properties of the language which are commonly included in works on credibility and deception detection [32,33,23]. We select the features most frequently deemed effective in predicting credibility: *number of words* in text and title, *average word length* in text and title, and *number of sentences* as well as *type-token-ratio* (TTR[3]) for the text.

External links describes the number of hyperlinks in the text which point to a different website. Linking to additional information, especially from external sources, has been found to increase credibility and be a predictor for legitimate news [19,33].

Readability of a text describes the complexity of its language in syntax and vocabulary, or the linguistic proficiency required to understand it. It is consistently featured in the literature as indicative of credibility or useful feature in fake news classification [19,33,30]. Textual web content with a higher reading level may be considered more professional, and therefore of higher quality, impacting credibility assessments. Using the

[3]TTR is the ratio of unique words (types) to total words (tokens) and describes a text's lexical density.

readability[4] library to compare various readability grades, we find that the Coleman-Liau index [34] performs best and use it in our signal evaluation.[5] The same readability index is also utilised by [35] and [33].

Emotional, sensational or affective language: Although previous results are at times inconsistent, researchers are in agreement overall that higher emotionality generally indicates content with lower credibility and an increased likelihood of being fake news [9,36,23]. Highly emotional messages aim at evoking specific feelings, sometimes in an attempt to hide poor argumentation or the absence of supporting evidence, while objective and unbiased information is commonly expected to be presented with little emotion. We decided to use the tool VADER for sentiment analysis and emotional word frequency analysis to target different aspects of language emotionality. For the latter we use the NRC Emotion Intensity Lexicon (NRC-EIL)[6] by [37] to calculate the average intensity per word with regard to eight different emotions.

Subjectivity or bias has been frequently named as a dimension of credibility [38,39,13], and was determined to be an advisable feature in automated credibility evaluation [17]. Like emotional language, this signal is difficult to determine precisely and consistently, nevertheless we opt to evaluate subjectivity due to its widely ascribed relevance to credibility assessments and low manipulability. We use TextBlob (through spaCy)[7] to capture a web page's subjectivity.

Punctuation: Previous research suggests that the analysis of punctuation, in particular question and exclamation mark usage may be beneficial for credibility assessments and fake news detection [32,10]. Increased question or exclamation mark presence – particularly in the headline – might indicate more sensationalist and clickbait'y content.

All caps are employed in informal communication, advertisements and the tabloid press to emphasise parts of text, and typically feel out of place and unprofessional in more formal contexts. Horne and Adalı [10] analysed fake news and determined an increased use of all caps compared to authentic news. Detection of these signals is trivial and they are thus highly interoperable, although it is important to exclude acronyms or initialisms as all-capitalised words.

Top-level domains: We analyse whether URLs include domains like "org", "gov" or "edu" as top- or second-level domain. Studies suggest that websites with these domains are assigned higher credibility ratings by consumers [25,26], which have been found to be a useful evaluation feature for credibility assessment systems [17,19]. While not content-related, this signal is interoperable, very easy to assess, and also difficult to game and therefore we decided to include it as an additional statistic on evaluated data.

Profanity: While insults, slurs or hate speech are discussed rarely in the web credibility literature, there are ample mentions of its negative effect on credibility in other domains [40]. For profanity, we merge a number of suitable word lists available online[8] and manually review the entries, removing those that are not necessarily insults or slurs

[4] https://github.com/andreasvc/readability/

[5] This index computes the readability based on on characters per word: $CLI = 0.0588L - 0.296S - 15.8$ where L is the average number of letters per 100 words, S is the average number of sentences per 100 words.

[6] https://saifmohammad.com/WebPages/AffectIntensity.htm

[7] https://spacy.io/universe/project/spacy-textblob

[8] https://github.com/dariusk/wordfilter/blob/master/lib/badwords.json,
https://www.freewebheaders.com/full-list-of-bad-words-banned-by-google/,
https://github.com/RobertJGabriel/Google-profanity-words,
http://www.bannedwordlist.com/lists/swearWords.txt

but perhaps only informal language, e. g., "hell", "crazy", "queer", "insane". Afterwards the text is searched for matching strings, which we find to occur extremely rarely.

A majority of the selected signals are specified precisely in terms of input information and evaluation manner, leading to high interoperability. For example, "misspellings in the text" is clearly defined, and can be consistently evaluated by different systems. On the other hand, a few signals are harder to specify precisely and their evaluation is more implementation-dependent (e. g., "emotionality of the language"). Some signals with lower interoperability, as well as some that are not on the list by [12], were nevertheless included after reviewing relevant publications in the literature, because of their low expected measurement difficulty and proven tie to credibility.

3.2. System Design

We combine pipelines for the measurement of a total of 23 credibility signals, grouped into ten evaluator modules, to assess a web page's credibility (Section 1). Beyond the general functionality of our system ALPACA (Automated Language-focused web page Credibility Assessor), our focus lies on the system's modularity and extensibility (see Figure 1). Subroutines for new signals can be added and modified without much effort, in order to be able to better study the effects of different signals and facilitate iterative improvements of the software.[9] Though a popular approach in automated credibility assessment, we do not train a machine learning model for the overall task. Our system is transparent, domain-independent, as well as modular and easily extensible in contrast to existing ML approaches, which can be considered black boxes.

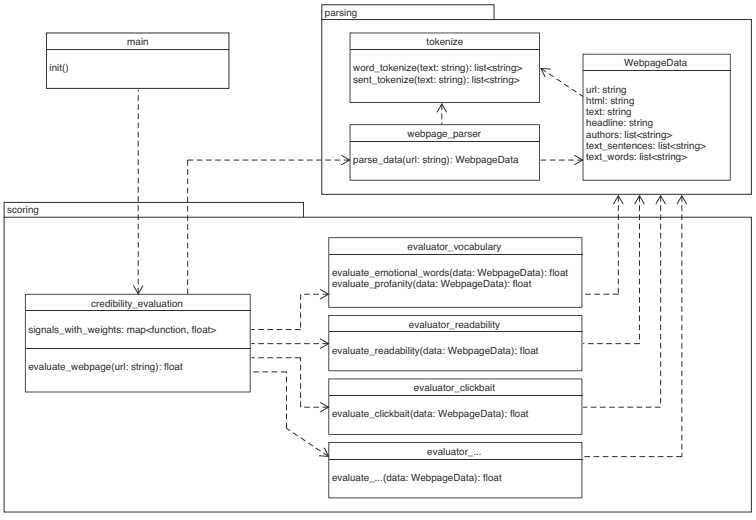

Figure 1. Simplified internal class diagram and architecture of our credibility evaluation system

We implement a web page parsing module and separate evaluation pipelines for the credibility signals, which process the content and return signal sub-scores. These are combined with signal weights through a linear combination function which empha-

[9]The data and code is available through our project's GitHub repository: https://github.com/lvap/alpaca.

sises low sub-scores to produce the final score between 0 (low) and 1 (high). We employ our own subroutines and a variety of publicly available tools in the signal evaluation pipelines. For several signals (emotional words, sentiment analysis, readability), we compared different approaches and selected the best-performing method; see Section 4 for the used data sets. Most language metrics are scaled by overall text length, i. e., number of words or sentences.[10]

Table 1. Summary of modules and credibility signals and weights to compute the score. +/– means the signal has a pos./neg. impact on the credibility score for greater values (e. g., more negativity in the headline leads to a lower score). * Domain ending is only weighted as bonus (if sub-score = 1). ** Profanity is only weighted as malus (if sub-score < 1).

Module	Credibility signal	Explanation	Weight
author	author	+ web page states author	0.1
clickbait	clickbait	– headline is clickbait	0.5
errors	errors	– grammar & spelling errors	0.35
language_structure	word_count_text	+ number of words in text	0.3
	word_count_title	– number of words in title	0.35
	sentence_count	+ number of sentences in text	0.4
	ttr	+ type-token-ratio	0
	word_length_text	+ average word length in text	0.3
	word_length_title	+ average word length in title	0.2
links	links_external	+ number of outbound links	0.1
readability	readability	+ Coleman-Liau grade	0.5
sentiment	polarity_text	+ polarity of text (pos. or neg.)	0.35
	polarity_title	– negativity of title	0.35
	subjectivity	– subjectivity	0.25
tonality	questions_text	– question marks in text	0.1
	questions_title	– title contains question mark	0.1
	exclamations_text	– exclamation marks in text	0.4
	exclamations_title	– title contains exclamation mark	0.25
	all_caps_text	– all caps words in text	0.1
	all_caps_title	– all caps words in title	0.4
url	domain_ending*	+ contains .org, .edu or .gov	0.5
vocabulary	profanity**	– no. of profanity words	0.1
	emotional_words	– avg. emotionality per word	0.7

The weights for the signals (Section 1) were determined iteratively, considering prior research results, signal measurement accuracy, and experimental calculations on test data. The type-token-ratio has no definite standing as credibility indicator in the literature and also did not correlate with credibility in our tests, thus it is not weighted in the final evaluation. We decided not to prioritise the domain weighting too much, as our

[10] In our repository we have dedicated methods for computing signal subscores, see https://github.com/lvap/alpaca/blob/signal-implementation-analysis/scoring/evaluator_url.py

system should still focus on content evaluation. Profanity only occurs extremely rarely, which is why we only treat it as malus.

We hypothesise that highly credible content scores well on most signals, while less credible web pages will have a mix of low and medium to high signal sub-scores. Therefore, we utilise a linear combination formula for the final score which increases the weight of low signal scores. For scores greater than 0.75, the original weight stays the same, but between 0.75 and 0.25 it linearly increases to twice the weight, and below 0.25 the score is constant again with doubled weight. For any signal α with sub-score $s_\alpha \in [0,1]$ and corresponding preliminary weight w_α from Section 1, the final weight towards the credibility score is:

$$\bar{w}_\alpha := w_\alpha(2 - min(max(2s_\alpha - 0.5, 0), 1)) \tag{1}$$

The system computes the signal sub-scores and final weights depending on the sub-scores and preliminary weights. It then performs a linear combination of the values to determine the overall credibility score between 0 (low credibility) and 1 (high).

4. Data Sets

We conduct test runs on a set of URLs compiled from several data sets with different domains to evaluate the credibility prediction performance of the entire system, of the individual credibility signals, and of the different signal implementations. Some of the data sets showed a variety of issues preventing their use as-is. We discuss the problems pertaining to the individual data sources in more detail in their respective sections. Three data sets contain credibility ratings on a Likert scale from one to five (least to most credible); the fourth contains pages classified as fake or real news; we interpret this as a credibility assertion, equating fake news with low and real news with high credibility.

The **Microsoft data set** [15] contains 1,000 URLs with corresponding Likert credibility ratings, assigned by one author. The authors performed 25 web searches (on 5 overall topics) and selected the top results as data set entries. Several early works in web credibility utilised this data to judge their systems' performance [17,18], but many of the URLs now redirect to unrelated content or point to nonexistent pages. While in theory cached versions of all pages were included, there are numerous practical problems: the cached version does not always match the actual URL, some pages are stored as HTML files and some as PDFs, the folder structure is inconsistent, and not all pages are contained [17,19]. A certain degree of domain and personal bias is expected due to the selection and rating methodology.

The **Reconcile data set** [41], also known as the Content Credibility Corpus (C3), spans 5,691 pages with almost 16,000 crowd-sourced evaluations by more than 2,000 annotators recruited through Amazon Mechanical Turk. The entries were selected manually through RSS feed subscriptions and Google queries by the authors, covering five major domains: politics & economy, medicine, healthy lifestyle, personal finance and entertainment. The annotators evaluated cached pages regarding several dimensions on a five-point Likert scale. Where necessary, we average the credibility ratings to obtain a mean page score. The data set is very diverse, such that we can assume tests on it to generalise well for the English-language web.

The **Credibility Coalition data set** [28] is a corpus of 42 web pages to enable research on credibility indicators. The entries were selected by finding the most shared articles on social media for specific search terms related to public health and climate science. These topics were chosen because the authors believe misinformation to be particularly prevalent in these domains despite an established expert consensus. Five experts for the domains were consulted to produce Likert-scale credibility ratings. On its own, the data set is modest in size, has a limited spectrum of topics, and the credibility ratings originate from just one person each (though they were produced by domain experts).

Lastly, we include the **FakeNewsNet data set** [42]. The fact-checking websites PolitiFact[11] GossipCop[12], automated Google searches and the platform E! Online[13], were utilised to obtain true and false articles in the domains of political news and celebrity gossip. A subset of the data set with roughly 17,400 real and 5,700 fake news pages is published. Although its size is substantial, and the fake news classifications sourced from established fact-checkers are convincing, the data is limited to just two domains.

We decide to use all suitable entries from the Credibility Coalition data set and 100 each (randomly selected) from the Microsoft, Reconcile, FakeNewsNet gossip and politics data sets respectively, for a total of 442 pages, which should yield a fairly robust corpus representative of the English-language web. Every URL was manually checked to confirm that it still links to the presumable original content. Especially older URLs point to unreachable web locations, highlighting the necessity to archive such data sets as HTML files or through online archival services. For broken links, we attempted to find cached versions in the Internet Archive[14]; if available, we chose the version with an archival date close to the release of the respective data set. Infrequently, an archived page instead of the original was selected to avoid cookie banners, pop-ups or redirects. Entries with content that clearly changes often and where the original state at the time of rating could not be determined – such as wiki-style websites, blog homepages, etc. – were discarded. Furthermore, satirical pages and those that are not text-centric, like image collections, link aggregations or video-focused pages, were also skipped.

For the evaluation, we merged the URLs into two data sets: one containing all web pages rated on a Likert scale, and another with the FakeNewsNet fake and real news. We refer to these collections as the "Likert" and "binary" data sets. The binary data contains 100 fake and 100 legitimate news pages. The Likert data set is rather unbalanced, encompassing (when rounding the credibility scores) 16 pages with rating 1, 24 with rating 2, 36 with rating 3, 109 with rating 4, and 57 with rating 5, of a total 242 pages. We process the data sets with our system to collect a variety of statistics pertaining to the different signals, the computed signal sub-scores and the overall web page credibility scores.

5. Results and Discussion

5.1. Individual Credibility Signals

We are particularly interested in the credibility prediction performance of individual signals. Emotional features, several morphological and syntactical text metrics, as well as

[11] https://www.politifact.com
[12] https://www.gossipcop.com, redirects to https://www.suggest.com as of 2022-01-21.
[13] https://www.eonline.com
[14] https://web.archive.org

Table 2. Summary data sets

Dataset	Description	#URLS used
Microsoft dataset [15]	1,000 URLs with corresponding Likert credibility ratings assigned by one author; contains websites with different topics and page types	100
Reconcile dataset [41]	5,691 thematically diverse pages with almost 16,000 crowdsourced evaluations by more than 2,000 participants	100
Cred. Coal. dataset [28]	43 web pages with most shared articles on social media related to climate science and public health rated by five domain experts	43
FakeNewsNet dataset [42]	Approx. 17,400 real news and 5,700 fake news in the domains of political news and celebrity gossip	200

the number of exclamation marks in the text and all capitals in the title best predict web content credibility. The domain ending, whether the headline is clickbait, and the sentiment of text and title are also influential. Many signals affirm our assumptions of their relation to credibility on at least one data set, however few produce correlations of statistical relevance for both (Section 3). The two signals that do are also among those with the highest ρ-coefficients: URL domain ending and emotional intensity of the text. The fact that the domain of a web page has such a clear link to credibility points to an influence of properties not related to the communicated information itself. The correlation values for emotional intensity confirm the importance of emotional signals for credibility assessment. We find that pages with more emotionally intense texts are less credible, while neutral or moderate language is linked to more credible content. Although sentiment analysis is rather average among all of our signals in terms of association with credibility, the polarity of the text and headline also show some trends. Specifically, less credible pages have texts and headlines with more negative sentiment on average, and the texts of more credible pages have a more positive sentiment.

Our experiments affirm that more credible web pages have generally longer content with more words and sentences, as well as longer words in the headline and text body. An exception is the length of the title, which is longer for less credible web pages. This coincides with the findings of Horne and Adalı [10], saying that fake news pages put the central claim and as much content as possible in the title, allowing users to skip reading the actual article. Horne and Adalı [10] also determine that fake news articles are repetitive and less complex, leading to a decreased type-token-ratio (TTR) in comparison to legitimate pages. However, we were unable to confirm this assertion, as the calculated TTR values on our data did not correlate positively with credibility. Some of the punctuation and all caps signals affected credibility, although the degree of association and its statistical relevance is small for most tonality signals. Exclamation marks in the text and all capitals in the headline are relatively strong in their correlation to the ratings on one data set respectively, while exclamation marks in the title have a small effect. This is not very surprising, as these features are indicative of more emotional, sensationalist and arguably unprofessional content. Frequency of question marks and all caps did not show a definite trend. Grammar and spelling errors have a negative influence on the credibility ratings of the Likert data set, but the relation on the binary data is not statistically distinctive enough ($p > 0.05$). It might be sensible to construct a more robust error checker if repeating the evaluations, as we suspect that a substantial part of the errors are false matches, contaminating the results. Clickbait is found to correlate with low credibility on the Likert data; the correlation coefficient is similar on the binary data but without

Table 3. Spearman correlation values for the data sets' web page credibility ratings and the underlying data used to compute the signal sub-scores (e. g., for errors the amount of errors per word, for readability the Coleman-Liau index), $*p < 0.05$

Signal	Likert data set		Binary data set	
	ρ	p	ρ	p
domain_ending	0.1746552141*	0.00645	0.2154101092*	0.00219
sentence_count	-0.0395744286	0.54008	0.3288565133*	0.00000
emotional_words	-0.155948275*	0.01517	-0.230712051*	0.00101
exclamations_text	-0.2768877057*	0.00001	-0.0584028802	0.41138
word_count_text	-0.0095320953	0.88272	0.3335141790*	0.00000
all_caps_title	-0.1219136512	0.06094	-0.1997281845*	0.00468
word_count_title	-0.0793427246	0.21972	-0.2014620755*	0.00423
errors	-0.1569651044*	0.01451	-0.1237231099	0.08091
clickbait	0.1362540513*	0.03451	0.1307937102	0.06489
polarity_text	0.035164685	0.58619	0.2280411934*	0.00116
polarity_title	-0.1564103010*	0.01508	-0.0764990174	0.28163
word_length_text	0.2519853282*	0.00007	-0.0055426319	0.93791
subjectivity	-0.1342034708*	0.03695	-0.0883356954	0.21355
readability	0.2375362725*	0.00019	-0.037585972	0.59722
exclamations_title	-0.1236448206	0.05525	-0.0467952685	0.51054
word_length_title	0.0679555930	0.29340	0.0765639770	0.28122
all_caps_text	-0.1059896996	0.09999	-0.0010784989	0.98791
questions_text	-0.0668347009	0.30045	0.0793350012	0.26412
author	0.0938657991	0.14543	-0.1286239389	0.06950
profanity	-0.0049520309	0.93891	0.0593909863	0.40350
questions_title	0.1467793630*	0.02266	-0.0368604890	0.60432
links_external	-0.2065647188*	0.00123	0.0072458950	0.91889
type-token-ratio	0.0229363492	0.72259	-0.3529098958*	0.00000

statistical significance. That clickbait content is less credible becomes apparent when analysing the distribution of clickbait headlines within our data sets: almost two-thirds of both the fake news and 1-rated web pages have titles classified as clickbait, but only 49% of real news and just 30% of 5-rated pages (although 49% still seems like a high number for legitimate content). A weak link is established between increased subjectivity and lower credibility on the Likert data. As the relevance of bias and subjectivity to credibility assessments is emphasised by literary sources [13,17], we believe that the underwhelming overall performance of this signal is due to the simple pre-trained model we employed, and better results might be achievable with a more advanced assessor. Though utilised in many existing credibility classification systems, (Coleman-Liau) readability only predicted credibility on the Likert data. Further analysis reveals that the readability grade distribution is highly similar for both binary data set classes, but follows a trend for the Likert data. It is therefore possible that readability predicts general credibility, but is not able to discriminate between real and fake news content.

Overall, several well-performing credibility signals confirm previous findings regarding their link to credibility, while many further signals show a sound correlation with credibility on one data set but none on the other. This could point to biases in the data selection, but may as well illustrate stylistic differences in the contained web pages.

While we produce a number of results supported by evidence and consistent with the literature, for other signals we did not get statistically significant results such as question mark frequency, author, and outbound links. Although drawn from a plurality of sources, the small size of our data basis (442 pages) is a concern, and we note that our conclusions may not be generally valid. Repeating the evaluation on a sufficiently large and balanced data basis could provide additional insights.

We also analyse co-occurrences of signals, i. e., inter-signal correlations. We calculate the Spearman correlation values for all signal combinations and focus on results with $p < 0.05$. Four signal tuples exhibit particularly strong correlations with each other. Readability and average word length are positively correlated, just like the number of words and sentences. These associations are intuitive: longer words are more complex and lead to a higher readability level, and more demanding texts likely include longer words (the Coleman-Liau index considers letters per words); furthermore, a larger number of sentences obviously correlates with a higher word count, and vice versa. Additionally, TTR is negatively correlated to both the number of sentences and words, which is also easily explainable as longer texts have a much higher potential for lexical repetition.

5.2. System Performance

We analyse the system's credibility assessment performance on the Likert and binary data sets. The calculated scores cluster in the range between 0.45 and 0.7 and fall above a minimum of 0.1986 and below a maximum of 0.9510 (Section 2). The fact that high scores of almost 1 are reached, but barely any under 0.2, might confirm our hypothesis of higher sub-scores being relatively common even among less credible pages, which had motivated our scoring function's design. The distinguishing feature of pages with low credibility is a (small) number of low sub-scores, which should be assigned more weight to be able to affect the overall score. It seems the scoring function is not able to properly leverage the different sub-scores in order to spread the final scores evenly, and therefore a strong bias towards a certain range of values can be observed.

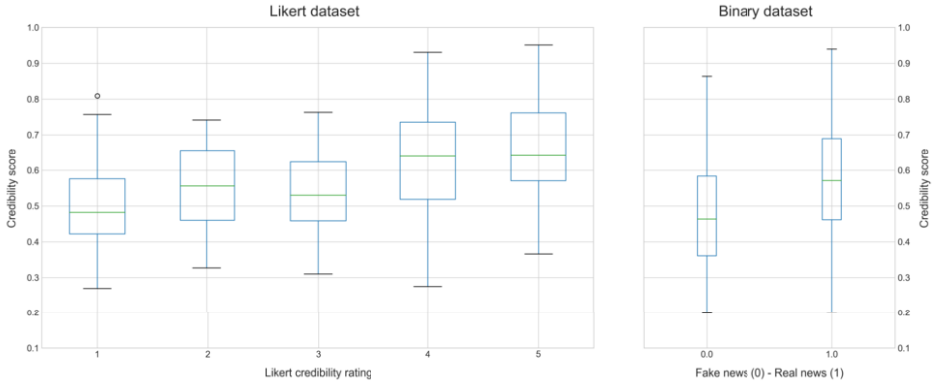

Figure 2. Distribution of computed credibility scores for Likert and binary data sets as box plots (Likert credibility ratings were rounded). The scores cluster in the range between 0.45 and 0.7.

The Spearman correlation coefficients for our credibility scores and the data sets' credibility ratings are $\rho = 0.29123$ (with $p = 0.00041$) for the Likert data and $\rho =$

0.30155 (with $p = 0.00001$) for the binary data. Our scores are undeniably correlated to the ratings, although the degree of correlation is quite weak given that the values should express the same information. The correlation strength is almost the same for the two data sets, suggesting a similar performance on both. To test the system as a credibility classifier, we assign classes to the web pages depending on the computed score. The binary data set's pages are classified as fake news if their credibility score is below 0.5, and as real news otherwise. We predict 94 pages to be fake news and 106 to be legitimate. 60 fake and 66 real news are labelled correctly for a total of 126 correct classifications, resulting in an accuracy of 63% (precision: 62.2%, recall: 66%). The accuracy is therefore above the random baseline but it leaves room for improvement. Web pages in the Likert data set are assigned five classes by multiplying their computed credibility score by five and rounding the resulting values. There are 2 pages with predicted class 1, 59 with class 2, 112 with class 3, 65 with class 4, and 4 with class 5. Comparing these classes to the rounded Likert ratings, we find that 68 out of 242 pages are labelled correctly for an accuracy of 28.1%. The mean offset between a web page's Likert rating and our computed score times five is 1.01828, and the median offset is 0.90077. Thus, our system's credibility scores differ by about 1 on average from the original Likert ratings. The classification performance is clearly better than a random baseline of 20%, but not completely convincing overall. Despite the Likert ratings being very imbalanced, our computed scores roughly follow a normal distribution around the centre class 3, which might be desirable for the general evaluation of web pages. Our system outperforms the random baselines for classifying web page credibility on a Likert scale or as fake/real news. This demonstrates that content-driven credibility evaluation is feasible, and that the corresponding signals are important components of web credibility. The system's performance can still be improved though, such as by adding further signals or improving the credibility score function. Statistics from the evaluation of a large data set could be used to devise an optimal formula for combining sub-scores as well as optimal signal weights.

5.3. Limitations and Future work

Many signals show a correlation with credibility on one dataset but none on the other. This may point to biases in data selection, but may as well illustrate stylistic differences in the contained web pages, e. g., higher readability grades (for all examined readability metrics) were linked to increased credibility on the Likert data, but were not associated with the fake or real news classes. Repeating the evaluation on a sufficiently large and balanced data basis could shed light on the root cause of this behaviour. The overall system outperforms the random baseline for both classifying fake news and web pages into credibility ratings on a Likert scale. This proves that content-focused credibility evaluation is feasible, and that content-related signals are important components of web credibility. However, the overall performance of the system is still in need of improvement.

Besides the addition of further signals, optimisation of the system should focus on the credibility score function and the measurement of individual signals. Theoretically, signal statistics from the evaluation of a large dataset could be used to devise an optimal formula for combining signal sub-scores, including optimal signal weights. We utilised the same datasets to determine the signal weights (through an analysis of the assessment data, together with other factors) and to test the system's performance; ideally, the system should be tested on unrelated data, and our approach could certainly lead to some-

what biased results. Regarding the signal measurements, in general, signal data is often very erratic, with many outliers that introduce unnecessary noise. A common reason for this might be parsing errors or flaws in the signal evaluation, such as in the grammar and spelling error pipeline, which frequently flags passages which are not true errors, or the subjectivity assessment, where we employ a simple pre-trained model, but which could be improved through the use of a better bias and subjectivity detector. Ultimately, while we produce a number of results supported by convincing evidence and consistent with scientific literature, the interpretations should be taken with a grain of salt. While drawn from a plurality of sources, the small size of our data basis (442 pages) is a concern and we must be aware that conclusions drawn from its evaluation may not be generally valid. Therefore, signal-credibility associations (or lack thereof) may not necessarily point to the actual performance of the signals, but could be induced by implementation or dataset specifics for the signals with ambiguous correlation to credibility according to our evaluation. These relationships remain to be further investigated by future research.

6. Summary and Conclusions

The credibility of web content has become an increasingly important public issue as communication and information exchange keeps evolving in the digital age. We believe that computational linguistics and artificial intelligence can play an important role in the development of technologies that help shift the online ecosystem towards more credible content through automated content evaluation and curation tools. When designing such technologies, we must take special care not to devise systems which facilitate censorship and foster social division or echo chambers. We present a system that automatically evaluates the credibility of web pages and produces a web page credibility score, to be utilised by users to inform and support their own assessments. We use the extensive credibility signal list published by the W3C Credible Web Community Group and results from previous research to identify key signals to include in our system. Our focus lies on properties that are linked to the actual content. We evaluate our approach on two data sets, each compiled by selecting a subset of random entries from several publicly available data sets to minimise possible biases. The first contains web pages and their credibility ratings on a five-point Likert scale, while the other consists of real and fake news articles, where the real news are assumed to have high credibility and the fake news low credibility. We can confirm previous findings that link credibility to several signals related to emotion, structural properties of the language, and punctuation and typesetting characteristics. Web pages with lower credibility have a greater emotionality, less complex language, are shorter, and contain more errors and exclamation marks. Headlines of less credible web content are longer, contain more words in all capitals, have more negative sentiment and are frequently clickbait: two-thirds of the web pages with low credibility in our evaluation data sets had clickbait titles. Low quality pages have shorter, less sophisticated content and put more emphasis on the headline, perhaps to allow readers to obtain all necessary information from the title and be able to skip reading the actual text. We use the computed credibility scores to assess the performance of our system, which achieves a classification accuracy of 63% for fake news detection, and a prediction accuracy of 28% for assigning credibility ratings on a five-point scale. Our system outperforms the random baselines of 50% and 20% respectively for both data

sets, affirming that content-focused credibility evaluation is feasible. In terms of future work, we intend to incorporate additional signal pipelines (e. g., inter-titles, number and length of paragraphs, photos) and improve the implementation of some existing signals which appear to not yet perform as anticipated. We plan to replicate the evaluation of the signals and the system as a whole on a larger, more robust and balanced data set to obtain more conclusive results. Lastly, some useful signals we analysed – number of sentences, sentiment of text and title, words in all capitals, length of the headline – were not directly covered by the W3C list of credibility signals, but we suggest to include them in any upcoming compilations of credibility signals based on our experimental results.

Acknowledgements

The research presented in this article is partially funded by the German Federal Ministry of Education and Research (BMBF) through the projects QURATOR (Unternehmen Region, Wachstumskern, no. 03WKDA1A) and PANQURA (no. 03COV03E).

References

[1] Zannettou S, Sirivianos M, Blackburn J, Kourtellis N. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. Journal of Data and Information Quality. 2019;11(3):1-37.

[2] Watson A. Share of adults worldwide who believe fake news is prevalent in selected media sources as of February 2019; 2021. Accessed 2022-01-21. URL https://www.statista.com/statistics/1112026/fake-news-prevalence-attitudes-worldwide/.

[3] Nakov P, Mihaylova T, Màrquez L, Shiroya Y, Koychev I. Do Not Trust the Trolls: Predicting Credibility in Community Question Answering Forums. In: Proc. of the Int. Conf. Recent Advances in Natural Language Processing, RANLP 2017. Varna, Bulgaria: INCOMA Ltd.; 2017. p. 551-60.

[4] Gupta A, Kumaraguru P, Castillo C, Meier P. In: Aiello LM, McFarland D, editors. TweetCred: Real-Time Credibility Assessment of Content on Twitter. Cham: Springer; 2014. p. 228-43.

[5] Chen Y, Conroy NK, Rubin VL. News in an online world: The need for an "automatic crap detector". Proc of the Association for Information Science and Technology. 2015;52(1):1-4.

[6] W3C Community Group Credible Web. Technological Approaches to Improving Credibility Assessment on the Web; 2018. Accessed 2022-01-21. Sandro Hawke, editor. URL https://www.w3.org/2018/10/credibility-tech/.

[7] Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, et al. The science of fake news. Science. 2018;359(6380):1094-6.

[8] Rehm G. An Infrastructure for Empowering Internet Users to Handle Fake News and Other Online Media Phenomena. In: Rehm G, Declerck T, editors. Language Technologies for the Challenges of the Digital Age. GSCL 2017. Springer; 2018. p. 216-31.

[9] Giachanou A, Rosso P, Crestani F. Leveraging Emotional Signals for Credibility Detection. In: Proc. of the 42nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. SIGIR'19. Association for Computing Machinery; 2019. p. 877-80.

[10] Horne B, Adalı S. This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News. Proc of the Int AAAI Conf on Web and Social Media. 2017;11(1):759-66.

[11] Karimi H, Tang J. Learning Hierarchical Discourse-level Structure for Fake News Detection. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics; 2019. p. 3432-42.

[12] W3C Community Group Credible Web. Credibility Signals; 2019. Accessed 2022-01-21. Sandro Hawke, editor. URL https://credweb.org/signals-20191126.

[13] Fogg BJ, Soohoo C, Danielson DR, Marable L, Stanford J, Tauber ER. How Do Users Evaluate the Credibility of Web Sites? A Study with over 2,500 Participants. In: Proc. of the 2003 conference on Designing for user experiences. DUX '03. Association for Computing Machinery; 2003. p. 1-15.

[14] Rieh SY, Belkin NJ. Understanding Judgment of Information Quality and Cognitive Authority in the WWW. In: Preston CM, editor. Proc. of the 61st Annual Meeting of the American Society for Information Science (ASIS). vol. 35; 1998. p. 279-89.

[15] Schwarz J, Morris MR. Augmenting Web Pages and Search Results to Support Credibility Assessment. In: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems. CHI '11. Association for Computing Machinery; 2011. p. 1245-54.

[16] Horne BD, Dron W, Khedr S, Adalı S. Assessing the News Landscape: A Multi-Module Toolkit for Evaluating the Credibility of News. In: Companion Proc. of the The Web Conf. 2018. WWW '18. Int. World Wide Web Conferences Steering Committee; 2018. p. 235-8.

[17] Olteanu A, Peshterliev S, Liu X, Aberer K. Web Credibility: Features Exploration and Credibility Prediction. In: Serdyukov P, Braslavski P, Kuznetsov SO, Kamps J, Rüger S, Agichtein E, et al., editors. Advances in Information Retrieval. ECIR 2013. Springer Berlin Heidelberg; 2013. p. 557-68.

[18] Wawer A, Nielek R, Wierzbicki A. Predicting Webpage Credibility Using Linguistic Features. In: Proc. of the 23rd Int. Conf. on World Wide Web. WWW '14 Companion. Association for Computing Machinery; 2014. p. 1135-40.

[19] Esteves D, Reddy AJ, Chawla P, Lehmann J. Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web. In: Proc. of the First Workshop on Fact Extraction and VERification (FEVER). Association for Computational Linguistics; 2018. p. 50-9.

[20] Afroz S, Brennan M, Greenstadt R. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In: 2012 IEEE Symposium on Security and Privacy. IEEE; 2012. p. 461-75.

[21] Rashkin H, Choi E, Jang JY, Volkova S, Choi Y. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2017. p. 2931-7.

[22] O'Brien N, Latessa S, Evangelopoulos G, Boix X. The Language of Fake News: Opening the Black-Box of Deep Learning Based Detectors; 2018. Paper presented at the Workshop on "AI for Social Good", NIPS 2018.

[23] Przybyła P. Capturing the Style of Fake News. Proc of the AAAI Conf on Artificial Intelligence. 2020;34(1):490-7.

[24] Keshavarz H. Assessing the credibility of Web information by university students: Findings from a case study in Iran. Global Knowledge, Memory and Communication. 2020;69(8/9):681-96.

[25] Rieh SY, Belkin NJ. Interaction on the Web: Scholars' judgment of information quality and cognitive authority. In: Kraft DH, editor. Proc. of the 63rd Annual Meeting of the American Society for Information Science (ASIS). vol. 37; 2000. p. 25-38.

[26] Fogg BJ, Marshall J, Laraki O, Osipovich A, Varma C, Fang N, et al. What Makes Web Sites Credible? A Report on a Large Quantitative Study. In: Proc. of the SIGCHI Conf. on Human Factors in Computing Systems. CHI '01. Association for Computing Machinery; 2001. p. 61-8.

[27] Hong T. The influence of structural and message features on Web site credibility. Journal of the American Society for Information Science and Technology. 2006;57(1):114-27.

[28] Zhang AX, Ranganathan A, Metz SE, Appling S, Sehat CM, Gilmore N, et al. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In: Companion Proc. of the The Web Conf. 2018. WWW '18. Int. World Wide Web Conferences Steering Committee; 2018. p. 603-12.

[29] Bourgonje P, Moreno Schneider J, Rehm G. From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. In: Proc. of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism. Association for Computational Linguistics; 2017. p. 84-9.

[30] Karadzhov G, Gencheva P, Nakov P, Koychev I. We Built a Fake News / Click Bait Filter: What Happened Next Will Blow Your Mind! In: Proc. of the Int. Conf. Recent Advances in Natural Language Processing. RANLP 2017. INCOMA Ltd.; 2017. p. 334-43.

[31] Beede P, Mulnix MW. Grammar, spelling error rates persist in digital news. Newspaper Research Journal. 2017;38(3):316-27.

[32] Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R. Automatic Detection of Fake News. In: Proc. of the 27th Int. Conf. on Computational Linguistics. Association for Computational Linguistics; 2018. p. 3391-401.

[33] Castelo S, Almeida T, Elghafari A, Santos A, Pham K, Nakamura E, et al. A Topic-Agnostic Approach for Identifying Fake News Pages. In: Companion Proc. of The 2019 World Wide Web Conference. WWW '19. Association for Computing Machinery; 2019. p. 975-80.

[34] Coleman M, Liau TL. A computer readability formula designed for machine scoring. Journal of Applied Psychology. 1975;60(2):283-4.

[35] Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B. A Stylometric Inquiry into Hyperpartisan and Fake News. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2018. p. 231-40.

[36] Li Q. Clickbait and emotional language in fake news; 2019. Accessed 2022-01-21. Preprint. URL https://www.ischool.utexas.edu/~ml/papers/li2019-thesis.pdf.

[37] Mohammad SM. Word Affect Intensities. In: Proc. of the Eleventh Int. Conf. on Language Resources and Evaluation. LREC 2018. European Language Resources Association (ELRA); 2018. .

[38] Metzger MJ, Flanagin AJ, Eyal K, Lemus DR, Mccann RM. Credibility for the 21st Century: Integrating Perspectives on Source, Message, and Media Credibility in the Contemporary Media Environment. Annals of the Int Communication Association. 2003;27(1):293-335.

[39] Rieh SY, Danielson DR. Credibility: A multidisciplinary framework. Annual Review of Information Science and Technology. 2007;41(1):307-64.

[40] Jay T. Cursing in America: A psycholinguistic study of dirty language in the courts, in the movies, in the schoolyards and on the streets. John Benjamins; 1992.

[41] Kakol M, Nielek R, Wierzbicki A. Understanding and predicting Web content credibility using the Content Credibility Corpus. Information Processing & Management. 2017;53(5):1043-61.

[42] Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media. Big Data. 2020;8(3):171-88.