Automatic Assessment of Online Content Credibility by Measuring the Adherence to Journalistic Standards

Eliza Danila[1], Julian Moreno-Schneider[2] and Georg Rehm[2]

[1] *TU Berlin, Berlin, Germany*
[2] *DFKI GmbH, Berlin, Germany*

Abstract

This paper analyses the correlation between online news articles' credibility and compliance with journalistic standards and ethics. Our main hypothesis is that news sources with low credibility adhere less to journalistic standards and ethics than high credibility sources. Additionally, the influence of web page layout factors is analysed. Our definition of metrics is based on the existing literature regarding journalistic best practices, as well as the list of credibility signals provided by the W3C Credible Web Community Group. News article credibility is assessed based on a score, which is calculated as a sum of linear subscores for the identified metrics divided in four credibility signal categories: Formality, Neutrality, Transparency and Layout. A curated dataset of 250 recent news items from known fake news sources, as well as 200 news items from established real news sources, is used in the testing phase. Although the comparison between real news and false news shows that, on average, the credibility score of genuine news is only 5% higher than that of fake news, the results do show more significant differences for certain subcategories and signals.

Keywords

Credibility assessment, credibility signals, fake news, journalism, journalistic standards, NLP

1. Introduction

In recent years, the term "fake news" has become ubiquitous, it is often misused for discrediting political opponents, yet targeted misinformation campaigns are still a reality. With the help of social media, fake news and "alternative facts" disperse at a rapid pace and reach a wide audience, which reduces social cohesion and leads to an erosion of trust in institutions, as argued by [1]. A recent poll among US citizens [2] revealed that the majority of respondents are very concerned about the spread of misinformation, especially if it is related to the coronavirus and vaccines. Also, the majority of respondents believes that information dispersed on social media is not accurate and is being tampered with by international state-actors. Nevertheless, at least one third of the respondents believes in theories related to the "deep state", satanic elite conspiracies or the creation of the coronavirus in a Chinese lab.

The increase in organized misinformation campaigns [3] and the seeming inability of a large number of news consumers to discern fake from real news motivated several countermeasures

Qurator 2022: 3rd Conference on Digital Curation Technologies, September 19-23, 2022, Berlin, Germany
danilaeliza@gmail.com (E. Danila); julian.moreno_schneider@dfki.de (J. Moreno-Schneider); georg.rehm@dfki.de (G. Rehm)
0000-0001-7116-9338 (J. Moreno-Schneider); 0000-0001-7116-9338 (G. Rehm)

CEUR Workshop Proceedings (CEUR-WS.org)

against misinformation. Facebook and Twitter introduced warnings for posts which contain false information on coronavirus pandemic-related topics as mentioned by Brown [4] and BBC News [5]. Additionally, the platforms provide links to verified resources regarding the coronavirus in order to facilitate access to genuine information. Other countermeasures include media literacy projects [6], fact-checking by independent organizations[1,2] or commercial offers for website credibility ratings comprised by analysts.[3] However, all these measures are costly, since the assessment is based on human analysis and requires a repetition of the process for each piece of information. Automated credibility assessment, on the other hand, allows for the analysis of more news items in shorter time. One example for automated credibility assessment is the US-based company The Factual, which uses four metrics to automatically assess the probability of news items being truthful and to provide them with a credibility score.[4]

The Credible Web Community Group of the World Wide Web Consortium (W3C) proposed 160 signals which can be used in credibility automation.[5] Since The Factual only seems to use four of these signals and does not disclose the exact method of calculating the credibility score, there is a need for an extension and refinement of the approach by experimenting with different combinations of credibility metrics. We propose an evaluation of the credibility of news items through a combination of metrics, based on the W3C signals.

In terms of its main contributions, this paper presents an automatic system that helps users to determine the credibility of online content by identifying signals which can be transformed into measurable metrics. We develop a credibility score which is viable for assessing a large number of diverse online news sources, focusing upon online news outlets, which excludes microblogs such as Twitter or other social media posts. We concentrate on online news items' compliance with established journalistic standards, ethics and best practices [7, 8]. The assessment aims at being self-contained (using only the content of the respective web page); it does not take into consideration information about its source or author, offering a neutral perspective about whether the information at hand *appears to be* credible. We take into account standards regarding layout aspects of a web page, which do not seem to have been leveraged in an automated manner in existing research. We also investigate the usage of tools typically used for automated software testing, leveraging their features for credibility testing, instead of using typical machine learning or decision theory approaches [9].

The remainder of this article is structured as follows. Section 2 describes the selection of signals. Section 3 explains the system architecture. Section 4 presents our results, the dataset used for testing and the most notable statistics. Potential shortcomings of the implementation, as well as an interpretation of results and possible approaches for future work, are detailed in Section 5.

[1] https://www.snopes.com

[2] https://www.mimikama.at

[3] https://www.newsguardtech.com/how-it-works/

[4] https://www.thefactual.com/how-it-works

[5] https://credweb.org/signals-20191126

2. Methodology

This section describes the methodology according to which the identified signals are transformed into metrics and combined into a final credibility score.

2.1. Selection and Measurement of Signals

The selection of signals for credibility assessment must take into consideration that (i) the measurement result must be reproducible at different points in time and by different persons or machines; (ii) the selected signals or combination of signals must be proven to correlate with content credibility; (iii) producers of content with low credibility should encounter high costs when trying to fake the signals used in a credibility measurement; and (iv) the signals must be clearly and unambiguously defined. Based on these constraints various signals have been identified for four categories (Layout, Formality, Neutrality, Transparency). Each signal is being attributed a score from 0 to 1. A detailed description of the selection criteria for each signal can be found in Danila [10].

- **Layout signals** [11]: **Number of pictures**, which is being put into relation with the article length; **Presence of a video**; **Font size** of both the headline and the text; and **Font type**.
- **Formality signals**: **Relative number of unique spelling errors** [12]; **Number of question and exclamation marks** in the title [13] and in the text [14]; number of consecutive exclamation or question marks[6]; **Presence of words in all capitals**, excluding headlines and acronyms [15]; and **Vocabulary size of the article** [16]: syntactic complexity, average sentence length, semantic complexity, a simple type/token ratio.
- **Neutrality signals**: **Number of superlatives** [17]; **Number of profanities**; **Racial slurs** [18]; and **Number of words with emotional valence** [19].
- **Transparency signals**: **Number of links** and **number of direct quotes** are aggregated and a linear score calculated [20]; **Number of external links** in relation to the number of links [21]; **Number of broken links** [21]; **Presence of an author**; and Whether an article is **marked as opinion** or not.

For the score components that are set in relation to the total number of words, a minimum acceptable ratio has been determined. The values are re-scaled based on a minimum value to better represent the difference between different news items' scores. The minimum value for most score components in question is 0.8. This leads to any article with spelling mistakes, non-neutral language, all capitalized words or unacceptable punctuation marks in a ratio of more than 20% achieving a score of 0 for this component. One exception is language complexity, for which a minimum value of 0.2 and a maximum value of 0.45 have been determined, since almost no articles fall outside this range.

[6] https://credweb.org/signals-20191126#signal-number-of-exclamation-points

2.2. Development of Credibility Score

The credibility score is comprised of four sub-categories: Layout, Formality, Neutrality and Transparency. Each category consists of subscores for the chosen signals pertaining to it. The final score is computed based on the Ordered Weighted Averaging method [22], which can be applied in model-driven credibility assessment and is a manner of solving Multi-Criteria Decision Making (MCDM) problems. The simple, unweighted credibility score can be formalized as follows:

$$\text{scoreCred} = \text{average}(\sum \text{score}_i) \tag{1}$$

where $i \in \{\text{Layout}, \text{Formality}, \text{Neutrality}, \text{Transparency}\}$ and

$$\text{score}_i = \text{average}(\sum \text{score}_i^j) \tag{2}$$

wherein the subscores (score_i^j) are calculated based on the average scores of the combined signals used to measure them, which lay in the interval [0, 1]

$$j \in \{\text{Pictures}, \text{Video}, \text{TextSize}, \text{TextType}\}^{i=\text{Layout}} \tag{3}$$

$$j \in \{\text{Spelling}, \text{Punctuation}, \text{Capitalized}, \text{Vocabulary}\}^{i=\text{Formality}} \tag{4}$$

$$j \in \{\text{Superlatives}, \text{Profanities}, \text{Slurs}, \text{Emotional}\}^{i=\text{Neutrality}} \tag{5}$$

$$j \in \{\text{Links}, \text{ExternalLinks}, \text{BrokenLinks}, \text{Author}\}^{i=\text{Transparency}} \tag{6}$$

Additionally, the Ordered Weighted Average allows for weights to be attributed to the different components of the score. In this paper, the score is first computed without weights and based on the observed differences (a deep description can be found in Danila [10]), the following weights are attributed to the credibility score components:

- Formality – Spelling: 0.2
- Formality – Punctuation: 0.3
- Formality – Language Complexity: 0.3
- Neutrality – Superlatives: 0.2
- Neutrality – Emotional words: 0.2
- Transparency – Citations: 0.2
- Transparency – Author mentioned: 0.2
- Layout – Video: 0.1 if video is present, else 0.0
- Layout – Font size and font type: 0.05 each

All other signals are included with a weight of 0.1.

3. Architecture and Implementation

This section provides a description of the credibility score automation process and implementation choices. The code is available via GitHub.[7] The code implemented by ourselves is open-source, although the concrete licenses of third party tools must be considered.

[7] https://github.com/Lzdnl/credibilityScore

Figure 1 shows the high-level architecture of the system. Using a URL provided by the user, the tool retrieves the relevant elements of a web page, performs cleaning actions and computes the corresponding metrics for each credibility category. Finally, based on the obtained metrics, a credibility score is computed.

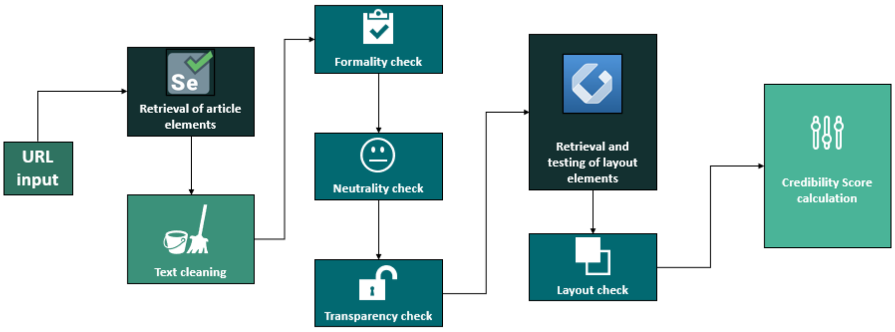
URL input
Retrieval of article elements
Text cleaning
Formality check
Neutrality check
Transparency check
Retrieval and testing of layout elements
Layout check
Credibility Score calculation

Figure 1: High-level architecture of our prototype system

Below we provide descriptions of the system components, following the order in which they are executed.

3.1. Fetching Elements from a Web Page

The elements of an article are retrieved from its web page using Selenium,[8] a framework for the automated interaction with browsers, which is widely used for website and web page testing. Selenium can recognize elements, which have been defined by the user based on HTML properties such as CSS selectors, IDs or XPath.

Before extracting elements from the website, a cookie banner has to be accepted due to GDPR [23]. Therefore, we installed the free-to-use Selenium third-party extension “I don’t care about cookies”[9], which automatically chooses the most easily feasible option for cookie banners.

Using Selenium, we obtain the following information: complete text extracted from the web page, article title, article URL, boolean value indicating the presence of an author and the list of links with additional information such as their text and the surrounding text.

3.2. Cleaning the Article Text

The second component of the system cleans the text obtained from the web page using a simple script in Selenium.[10] The first step in the text cleaning function is splitting the text obtained from the web page at the newline character and storing paragraphs in a list. Elements which are not part of the article text are removed (e. g., paragraphs which do not contain any punctuation marks or are very short, paragraphs containing expressions such as “cookie policy” or “all rights

[8] https://www.selenium.dev

[9] https://www.i-dont-care-about-cookies.eu

[10] Text extraction and cleaning can also be achieved with the aid of boilerplating tools such as Beautiful Soup, https://pypi.org/project/beautifulsoup4/.

reserved" etc.). This component returns the cleaned text, its total number of words and the total number of sentences.

3.3. Formality Check

The Formality component uses the article's cleaned text and title. It counts the number of exclamation and question marks found in the title, the number of overall consecutive question or exclamation marks found, as well as the number of question marks found in the text; there is no statistical difference in the usage of exclamation marks, see Yang et al. [14]. Additionally, words written in all capitalized characters are counted. To avoid false positives by counting abbreviations, only sequences of at least two consecutive words are considered.

Later, spelling mistakes are identified with the aid of the Python library "autocorrect".[11] Additionally, to enable users to pass their own credibility judgements based on the automated score, the system allows for displaying a list of misspelled words. The last step of the formality analysis is calculating the semantic complexity of the article, which is achieved by a simple type/token ratio – number of unique words divided by the total number of words.

The component returns the number of question or exclamation marks in the title, number of question marks in the text, overall number of consecutive question or exclamation marks, number of words for which all characters are capitalized, number of spelling mistakes in the title and in the text, lexical richness ratio, list of misspelled words, and list of the identified words for which all characters are capitalized.

3.4. Neutrality Check

The neutrality check analyses the content. It counts the instances of superlatives, which denote exaggerations, as well as words with emotional valence, profanities and racial slurs. Superlatives are counted by comparing the cleaned article text to a list of superlatives for more than 5400 English adjectives [24].[12]

The number of emotional words is calculated similarly. A lexicon containing words with emotional valence, as well as the type of emotion they convey, is used: EmoLex, the NRC Word-Emotion Association Lexicon[13] is a crowd-sourced lexicon [25]. To identify profanities, an unofficial list of words banned by Google is used[14].

Racial slurs are identified by comparing the cleaned text with entries from a racial slur database.[15] The list contains more than 2600 entries, however, many of them are homonyms of non-offensive words which can only be regarded as slurs in a certain context. To avoid false positives, the list has been cleaned; the modified list contains only those words which can be regarded as racial slurs regardless of context.

This component returns: number and list of identified superlatives, number and list of identified words with emotional valence, number and list of identified profanities and number and list of identified racial slurs.

[11] https://github.com/fsondej/autocorrect

[12] https://github.com/prosecconetwork/The-NOC-List/blob/master/NOC/DATA/TSV%20Lists/superlatives.txt

[13] https://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[14] https://www.freewebheaders.com/full-list-of-bad-words-banned-by-google/

[15] http://www.rsdb.org

3.5. Transparency Check

The Transparency category uses the list of links obtained previously, as well as the boolean value which indicates whether an author has been found. First, we delete links from buttons, related articles or advertisements. The list contains not only the links, but also their text and the surrounding text. The cleaning process deletes all links for which the surrounding text is not part of the cleaned article text. Empirical observation led to the conclusion that this method eliminates nearly all irrelevant links.

The remaining links are counted to obtain the number of references. For the number of external references, the domain of the article's URL is extracted and compared to that of each reference. The ones which do not match are counted as external links. The script identifies broken links by sending requests for the head of each reference URL and asserting that the response code is smaller than 400. In addition to the number of links, the number of direct quotes is calculated by counting all quotation marks and dividing them by 2. Finally, the script checks if the article is an opinion piece by identifying the word "opinion" in the URL.

The component returns the following elements: number of references, number of external references, number of broken links, number of direct quotes, list of URLs of all references, list of URLs of broken links and a boolean value indicating whether the article is marked as opinion.

3.6. Layout Check

Since Selenium is mainly designed for text-based assessments, the signals chosen for the credibility category Layout cannot be processed with this tool. Galen, however, provides a wide variety of features, being designed for layout tests in responsive web design.[16]

Galen provides a number of features which can be leveraged to measure the Layout metrics. Videos, pictures, the headline and the article text can be predefined as objects. However, the results provided in JSON must be further processed to obtain the desired metrics.

For videos, it suffices to assert that a video is visible on the web page. Extracting images is simple, as they are always displayed with the aid of an img tag. In the case of pictures, a simple count is not sufficient, because it is not possible to automatically differentiate between article pictures and other pictures, such as the website's logo or pictures from the related articles. Therefore, Galen checks the size of each picture and checks whether it has a width of at least 400 pixels, a measure identified by empirical testing, to avoid user avatars, logos and pictures of related articles. Galen can also extract font sizes and font types.

The component returns the following: number of pictures, presence of a video, information about whether the headline font size and text font size fall into the expected respective range, information about whether the headline font type and text font type are serif.

3.7. Calculating the Credibility Score

The final step is calculating the credibility score based on the collected measurements (see Section 2). The corresponding script performs simple operations for each credibility category

[16] http://galenframework.com

based on the data points retrieved from all previously executed components. Finally, it computes a weighted average based on the weights attributed to each metric.

4. Experiments

This section presents the results of the testing phase conducted for the automated credibility score calculation, describes the datasets used and compares the results. The Jupyter Notebook used for the data analysis is available in GitHub.[17]

4.1. Dataset

Although various datasets used for fake news detection have been published in recent years, identifying a collection of URLs for the purpose of testing our prototype was a challenge. Most fake news datasets focus on classifying news items into fake and real, i. e., they do not provide a score similar to ours. One such example is Shu [26], which offers two collections of fake and real news, one from the political domain, the other focused on celebrity gossip. Other datasets, like the one used by Zhang et al. [27], contain crowd-sourced credibility estimations on a five-point Likert scale. However, the criteria used for assessing the credibility of news items does not match our signals. Additionally, the dataset appears to be relatively limited. Yet another shortcoming relates to datasets only containing entries for fake news, without real news entries to compare against. One such example are the top 50 most popular fake news collected by BuzzFeed.[18]

Initially testing our system with existing datasets led to distorted results due to a number of reasons. For one, many of the URLs in the datasets do not exist anymore. The datasets were cleaned before testing by sending an automated request to the URL and removing the URLs from the list that result in an HTTP error code. However, not all websites generate HTTP error codes if a URL cannot be found. Several websites redirect to their homepage instead, rendering an automated removal of broken links infeasible. Additionally, some of the URLs in the existing datasets seem to be accessible only to registered users, which means that the URL redirects to either a login page or a subscription offer. Besides, a high number of entries were URLs for articles stored in the Internet Archive[19], in which case the automated credibility testing led to processing times of several minutes per URL. Also, due to the fact that some of the archived articles seem to have a modified HTML structure because of the archiving process, some page elements are not being properly recognized. Finally, not all datasets fit the scope of our work. While the system has been developed for online articles from newspapers in HTML format, some datasets also include Youtube videos, Twitter posts, PDF files and other content in formats that we consider out of scope. This results in partially very low credibility scores for these types of content.

Because of these shortcomings, a new dataset was collected. It consists of 250 entries for fake news and 200 entries for real news. The fake news URLs were collected by navigating to

[17] https://github.com/Lzdnl/credibilityScore

[18] https://github.com/BuzzFeedNews/2017-12-fake-news-top-50

[19] http://archive.org

websites known for spreading fake news, hate speech and conspiracy theories, a list of which had been compiled by the company Prophecy as part of a dataset for fake news detection.[20] Many of the domains listed in the collection were either expired or flagged as posing a security risk. Three URLs per domain were randomly collected from the domains which could still be accessed. The genuine news were collected from the websites of established English language newspapers and randomly choosing three URLs per domain. These newspapers are not limited to those from English-speaking countries, but also English editions of international newspapers. Additionally, both globally established and small local newspapers have been taken into account, resulting in a diverse dataset.

4.2. Comparison of Credibility Score for Fake and Real News

At first glance, the difference between the credibility scores for genuine and fake news does not seem significant. However, real news do perform better than fake news. Figure 2 shows that the mean credibility score for real news is higher by 4.9%, while the mean of the weighted score is higher by 6.3% in real news. Similar differences can be observed for the median. It is higher for real news, by 5.4% for the simple score and 6.5% for the weighted score.

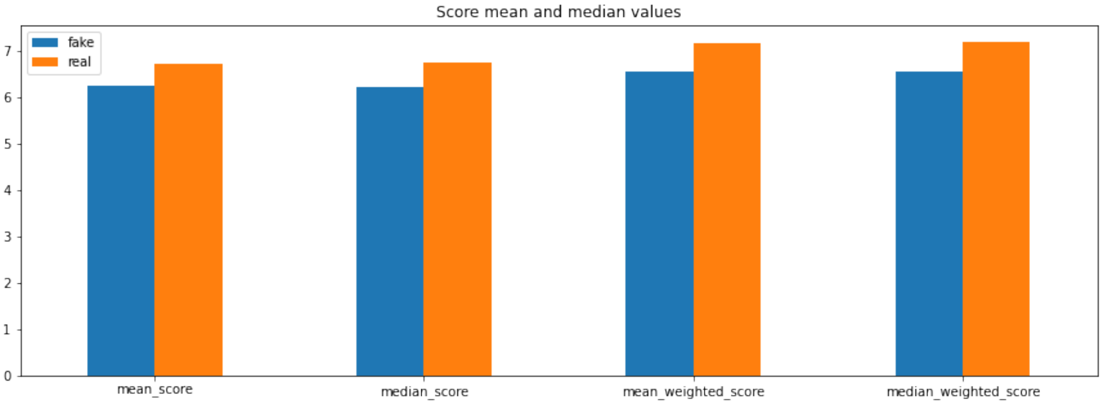

Figure 2: Mean and median comparison of simple and weighted score

When comparing the subscores for the categories Formality, Neutrality, Transparency and Layout, the same trend can be observed. As shown in Figure 3, on average, real news perform better. The highest difference, of 8.4%, can be observed in the Formality category, followed by Transparency with a difference of 5.7%, and Layout and Neutrality with a difference of 2.8% and 2.7%, respectively.

A deep dive into the components of the score categories reveals differences at component level. Figure 4 shows that most used signals lead to a higher score for real news. With a 22.4% difference, *punctuation* seems to be the strongest of the analysed indicators. The next indicator is the *font type* – usage of serif fonts – with a difference of 14.9%. Another well-performing layout indicator is *video*, with a difference of 10.5%. Two transparency indicators complete the

[20] https://github.com/several27/FakeNewsCorpus/blob/master/websites.csv

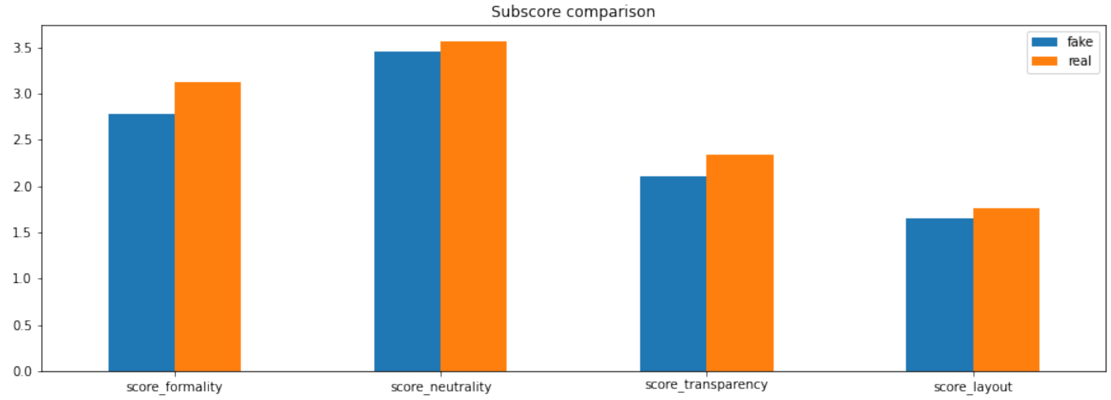

Figure 3: Mean comparison of subscores

top 5: the presence of *references* – as an aggregation between links and direct quotes – and the presence of *broken links*, with differences of 13.4% and 9.6%, respectively.

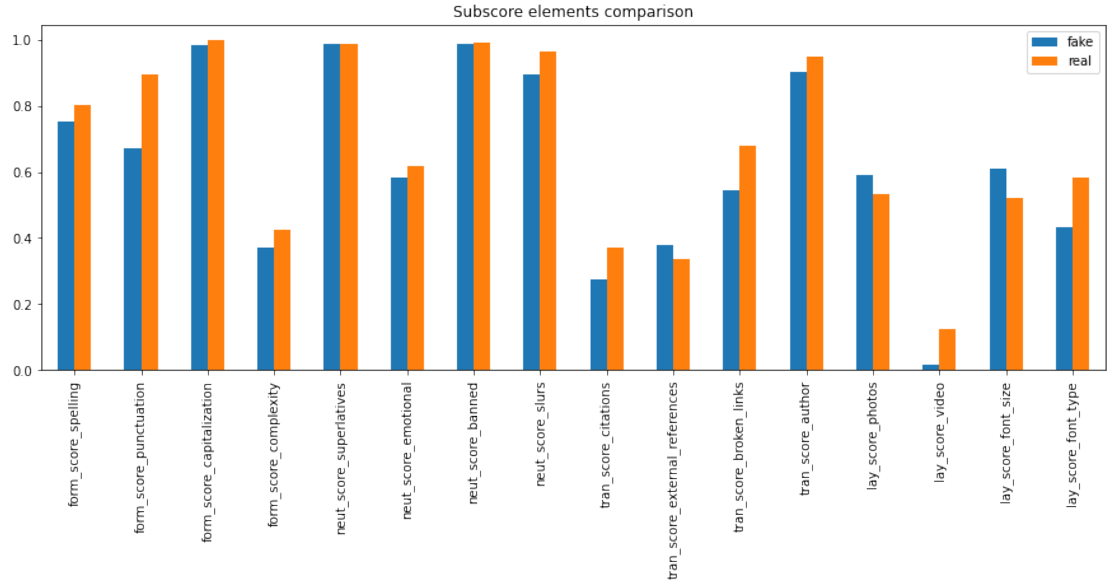

Figure 4: Mean comparison of subscore components

Notably, fake news perform better on three subscore components. The layout indicators *font size* and *pictures* show a difference in measurement of 8.8% and 5.6%, respectively, while the indicator *external references* shows a difference of 4.2%. This indicates that the font size might not be a reliable indicator. When it comes to pictures, it appears that most fake news articles are relatively short and contain one picture, which results in the maximum score for this indicator [11]. An evaluation of the pictures' originality could be taken into account for refining the score.

Finally, in the case of external references, it seems that credible articles are penalized for providing a high amount of links in general. They will therefore obtain a high score for references in general. However, the external reference score is calculated as a ratio between external links and all links. Therefore, articles with a lower amount of overall links can obtain a higher score for external references more easily.

Other notable differences can be observed by comparing the distribution of the credibility score (Figure 5):

- Scores below 5 are achieved by 7.8% of fake news items and 1.1% of real news items.
- Scores between 5 and 6 are achieved by 30% of fake news items and 16.4% of real news items.
- Scores between 6 and 7 are achieved by 41.3% of fake news items and 48% of real news items. Notably, it seems that in this case most of the real news are situated in the range between 6.5 and 7, while most of the fake news achieve scores in the range between 6 and 6.5.
- Scores between 7 and 8 are achieved by 20.4% of fake news items and 31% of real news items.
- Scores above 8 are achieved by 0.4% of fake news items and 3.5% of real news items.

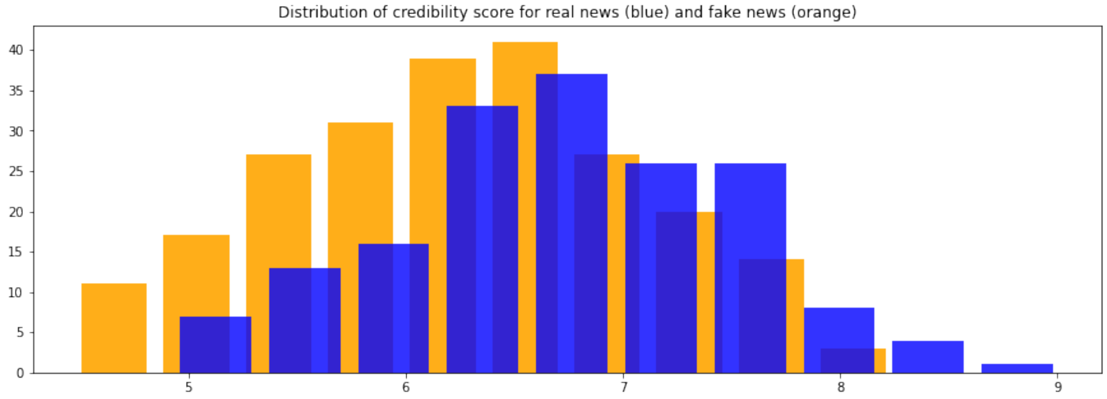

Figure 5: Distribution of credibility score for fake (orange) and real (blue) news

The weighted credibility score leads to a general shift towards higher scores in general (Figure 6). However, we notice bigger differences when it comes to the score distribution. With the simple score, 90.4% of fake news items and 95.4% of real news items achieve a score between 6 and 8. With the weighted score, 93.1% of fake news items achieve a score between 5 and 8, while 94.7% of real news items are situated between 6 and 9.

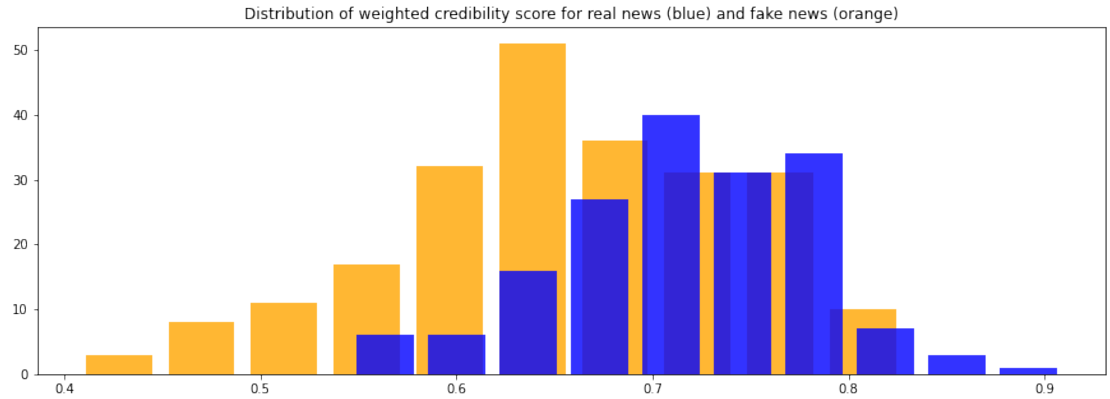

Figure 6: Distribution of weighted credibility score for fake (orange) and real (blue) news

4.3. Notable Examples

After the quantitative evaluation, the difference in the scores between real and fake news is not as significant as expected, which is why we also performed a qualitative evaluation of various notable examples.

For one, not all news from credible sources achieved high credibility scores. However, the low scores also stem from the fact that the news items in question do not fulfil the criteria chosen for measuring credibility. The most notable examples in this category are two articles from the Arabic newspaper Al Jazeera[21,22]. Both have a high Formality and Neutrality score. However, both articles achieve a low Transparency score, due to the low number of references and external links and the fact that the author is not mentioned. The lack of a video and the unbalanced number of pictures in relation to the length of the article account for an additional decrease in the Layout score.

Another surprising result is the high credibility score obtained by some content sources known for spreading fake news. The most notable example is an article from the conspiracy-focused website dcclothesline.com.[23] The article achieves a credibility score of 7.98. Although it was published on a website with very low credibility, the article fulfils the chosen credibility criteria to a high degree: low rate of spelling errors, does not use question or exclamation marks or all capitalized words and fulfils most of the established layout criteria. Additionally, it contains a high number of direct quotes and links, all of which are external, a low number of broken links, and has a named author.

A more precise credibility ranking of news sources can be achieved with an expanded dataset. Based on the limited dataset used for our evaluation, the top 5 real news domains with the highest mean credibility scores belong to: the US news channel MSNBC, the South African

[21] https://www.aljazeera.com/economy/2021/5/24/china-crackdown-forces-crypto-mining-operators-to-end-operations

[22] https://www.aljazeera.com/news/2021/5/24/coup-claim-as-samoas-elected-leader-locked-out-of-parliament

[23] https://www.dcclothesline.com/2021/05/23/total-tyranny-well-all-be-targeted-under-the-governments-new-precrime-program/

economy-focused newspaper businesslive.co.za, the English version of the German news source Deutsche Welle, the Guardian and the political news source Politico.

For fake news, the top 5 domains by simple credibility score are WorldNetDaily, the celebrity tabloid Upliftingtoday.com, the US right-wing platform Yesimright.com, the conservative Subjectpolitics.com, which publishes news and the news aggregator En.newsner.com.

5. Conclusions

We have identified a set of credibility signals, which can be measured in a reproducible and automated way, and clustered them into four credibility categories based on journalistic best practices and ethical principles. We have developed an optimizable credibility score function and applied the function to a dataset collected from known sources for both fake and real news. The implementation of the credibility assessment has been conducted with the aid of two automation tools for software testing: Selenium, which has been used for extracting content features of news items (e. g., headline, body text, references, author) and Galen, which has been leveraged for extracting layout features (e. g., font size and type, pictures and videos). The possibility for a precise definition of elements based on the HTML structure of the web page confirms that testing frameworks can indeed be leveraged for automated credibility assessment. However, there are some caveats. Firstly, precise definitions of elements guaranteed to function with a large number of websites are a laborious task which can only be achieved by testing a very large number of diverse websites. For this reason, further refinement of element definitions and cleaning of extracted elements is necessary.

Our results show a small difference between real and fake news items, with real news achieving higher scores on average. When regarding the results by category, the Formality and Transparency categories show higher differences than the Neutrality and Layout categories. A deeper dive into the analysed signals reveals some notable signals from the Formality, Transparency and Layout category, for which a relatively high difference can be observed: non-standard punctuation, the number of citations, the number of broken links, the presence of a video and the font type. Notably, the last two signals are exclusively layout-based. More research into how layout aspects influence credibility will be conducted in the future to further test this hypothesis.

In terms of future work, apart from the refinement of the definition of elements extracted from a web page, the optimization of the subscore scaling and the credibility score function are topics which might benefit from further research. Additionally, an expansion of the scope beyond surface credibility and a more complex definition of the developed categories would be valuable additions to the present work. More precise results might be obtained by regarding the proposed approach as one perspective of credibility assessment and integrating it into a holistic project which also takes into account other aspects and types of credibility.

Acknowledgments

The work presented in this article has received funding from the German Federal Ministry of Education and Research (BMBF) through the projects QURATOR (no. 03WKDA1A) and

PANQURA (no. 03COV03E) and from the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project SPEAKER (no. 01MK19011).

References

[1] A. Witze, How to detect, resist and counter the flood of fake news, 2021. URL: https://www.sciencenews.org/article/fake-news-misinformation-covid-vaccines-conspiracy.

[2] M. Newall, More than 1 in 3 americans believe a 'deep state' is working to undermine trump, 2020. URL: https://www.ipsos.com/en-us/news-polls/npr-misinformation-123020.

[3] L. von Richthofen, Covid disinformation campaign targeted biontech-pfizer, 2021. URL: https://www.dw.com/en/covid-disinformation-campaign-targeted-biontech-pfizer/a-57702440.

[4] A. Brown, Facebook will begin warning you if you interact with fake coronavirus news, 2020. URL: https://www.forbes.com/sites/abrambrown/2020/04/16/facebook-will-begin-warning-you-if-you-interact-with-fake-coronavirus-news/.

[5] BBC News, Coronavirus: Twitter will label covid-19 fake news, 2020. URL: https://www.bbc.com/news/technology-52632909.

[6] A. M. Guess, M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, N. Sircar, A digital media literacy intervention increases discernment between mainstream and false news in the united states and india, Proceedings of the National Academy of Sciences 117 (2020) 15536–15545. URL: https://www.pnas.org/content/117/27/15536.

[7] S. Cohen, J. T. Hamilton, F. Turner, Computational journalism, Commun. ACM 54 (2011) 66–71. URL: https://doi.org/10.1145/2001269.2001288. doi:10.1145/2001269.2001288.

[8] N. Diakopoulos, Automating the News: How Algorithms Are Rewriting the Media, Harvard University Press, 2019. URL: http://www.jstor.org/stable/j.ctv24w634d.

[9] S. Castelo, T. Almeida, A. Elghafari, A. Santos, K. Pham, E. Nakamura, J. Freire, A topic-agnostic approach for identifying fake news pages, in: Companion Proceedings of The 2019 World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 975–980. URL: https://doi.org/10.1145/3308560.3316739. doi:10.1145/3308560.3316739.

[10] E. Danila, Automatic Assessment of Online Content Credibility by Measuring the Adherence to Journalistic Standards, Master's thesis, Technische Universität Berlin, Berlin, Germany, 2021.

[11] J. O. Wobbrock, L. Hattatoglu, A. K. Hsu, M. A. Burger, M. J. Magee, The goldilocks zone: young adults' credibility perceptions of online news articles based on visual appearance, New Review of Hypermedia and Multimedia 0 (2021) 1–46. URL: https://doi.org/10.1080/13614568.2021.1889690.

[12] P. Beede, M. W. Mulnix, Grammar, spelling error rates persist in digital news, Newspaper Research Journal 38 (2017) 316–327. URL: https://doi.org/10.1177/0739532917722766.

[13] N. Stewen, Can any headline that ends in a question mark be answered by the word "no"?, 2020. URL: https://medium.com/@nicolos/can-any-headline-that-ends-in-a-question-mark-be-answered-by-the-word-no-6f93f6ce85e9.

[14] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, P. S. Yu, TI-CNN: convolutional neural networks for fake news detection, CoRR abs/1806.00749 (2018). URL: http://arxiv.org/abs/1806.00749.

[15] B. Horne, S. Adali, This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 11, AAAI Press, Montreal, Quebec, Canada, 2017, pp. 759–766.

[16] P. Tolochko, H. G. Boomgaarden, Analysis of linguistic complexity in professional and citizen media, Journalism Studies 19 (2018) 1786–1803. URL: https://doi.org/10.1080/1461670X.2017.1305285.

[17] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, Y. Choi, Truth of varying shades: Analyzing language in fake news and political fact-checking, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2931–2937. URL: https://www.aclweb.org/anthology/D17-1317.

[18] T. Riebe, K. Pätsch, M.-A. Kaufhold, C. Reuter, From conspiracies to insults: A case study of radicalisation in social media discourse, in: R. Dachselt, G. Weber (Eds.), Mensch und Computer 2018 - Workshopband, Gesellschaft für Informatik e.V., Bonn, 2018, pp. 595–603. URL: https://doi.org/10.18420/muc2018-ws12-0449.

[19] B. Ghanem, P. Rosso, F. Rangel, An emotional analysis of false information in social media and news articles, ACM Trans. Internet Technol. 20 (2020) 1–18. URL: https://doi.org/10.1145/3381750.

[20] M. Duncan, K. B. Culver, D. McLeod, C. Kremmer, Don't quote me: Effects of named, quoted, and partisan news sources, Journalism Practice 13 (2019) 1128–1146. URL: https://doi.org/10.1080/17512786.2019.1588148.

[21] W. Choi, B. Stvilia, Web credibility assessment: Conceptualization, operationalization, variability, and models, Journal of the Association for Information Science and Technology 66 (2015) 2399–2414. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23543.

[22] G. Pasi, M. D. Grandis, M. Viviani, Decision making over multiple criteria to assess news credibility in microblogging sites, in: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, Glasgow, United Kingdom, 2020, pp. 1–8. URL: https://doi.org/10.1109/FUZZ48607.2020.9177751.

[23] R. Schmitt, Auslegung der datenschutzrechtlichen Einwilligung – Anforderungen an Cookie-Banner konkretisiert, 2020. URL: https://blog.handelsblatt.com/rechtsboard/2020/06/12/auslegung-der-datenschutzrechtlichen-einwilligung-anforderungen-an-cookie-banner-konkretisiert/.

[24] T. Veale, Round up the usual suspects: Knowledge-based metaphor generation, in: Proceedings of the Fourth Workshop on Metaphor in NLP, Association for Computational Linguistics, San Diego, California, 2016, pp. 34–41. URL: https://www.aclweb.org/anthology/W16-1105.

[25] S. Mohammad, P. Turney, Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon, in: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics, Los Angeles, CA, 2010, pp. 26–34. URL:

https://www.aclweb.org/anthology/W10-0204.

[26] K. Shu, FakeNewsNet, 2019. URL: https://doi.org/10.7910/DVN/UEMMHS. doi:10.7910/DVN/UEMMHS.

[27] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins, E. Bice, S. Hawke, D. Karger, A. X. Mina, A structured response to misinformation: Defining and annotating credibility indicators in news articles, in: Companion Proceedings of the The Web Conference 2018, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 603–612. URL: https://doi.org/10.1145/3184558.3188731.