Reliability Prediction for Health-Related Content: A Replicability Study

Marcos Fernández-Pichel[1(✉)], David E. Losada[1], Juan C. Pichel[1], and David Elsweiler[2]

[1] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
{marcosfernandez.pichel,david.losada,juancarlos.pichel}@usc.es
[2] University of Regensburg, Regensburg, Germany
david@elsweiler.co.uk

Abstract. Determining reliability of online data is a challenge that has recently received increasing attention. In particular, unreliable health-related content has become pervasive during the COVID-19 pandemic. Previous research [37] has approached this problem with standard classification technology using a set of features that have included linguistic and external variables, among others. In this work, we aim to replicate parts of the study conducted by Sondhi and his colleagues using our own code, and make it available for the research community (https://github.com/MarcosFP97/Health-Rel). The performance obtained in this study is as strong as the one reported by the original authors. Moreover, their conclusions are also confirmed by our replicability study. We report on the challenges involved in replication, including that it was impossible to replicate the computation of some features (since some tools or services originally used are now outdated or unavailable). Finally, we also report on a generalisation effort made to evaluate our predictive technology over new datasets [20,35].

Keywords: Reliability · Language · Health-related content

1 Introduction

The emergence of digital media has brought a change in the way people inform themselves [33]. In many ways, this change has been positive, providing accessibility of information and speed of access, but we must also be aware of the dangers involved. The results offered can be unreliable [2], inaccurate [9], or of poor quality [34]. This can have a greater or lesser impact depending on the context [37], but is especially sensitive when it comes to **health-related content**, as Pogacar et al. [31] showed in a recent study.

Medical hoaxes, miracle diets, or advice given by unqualified people abound in this type of media [36] and can be highly dangerous if taken as true and

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 47–61, 2021.
https://doi.org/10.1007/978-3-030-72240-1_4

applied without the supervision of a medical professional. This has become particularly evident in the context of the pandemic we are facing, with substantial information about **COVID-19** being either dubious or of poor quality [19,30].

Often, **language** is a powerful indicator of the veracity of the contents [24]. Hidden patterns can be discovered not only by analysing the latent topics discussed in a certain text but also by studying the use of certain words [28]. An example is the use of **technical terms** or formalisms, which is usually associated with documents of higher quality and, in many cases, of greater reliability.

In this work, we report on our endeavours to replicate the predictive technology developed in Sondhi et al. [37], based on Natural Language Processing (NLP) and Machine Learning techniques. We chose this study since, to our knowledge, it was the first one to address the issue of automatically assessing the reliability of webpages in the medical domain. They reduced this problem to a binary-classification task. Moreover, they also provided a test dataset and a set of features to be taken into account (see Sect. 3).

If the results could be recreated, the conclusions extracted in the original study would be verified and reinforced. This replication effort is worthwhile to establish the utility of current technology, and its potential to be applied in filtering non-reliable content.

To this end, we examined and, where possible, re-implemented the features proposed by the original authors. In order for the results to be comparable, we applied the same experimental methodology and performance metrics proposed in the original paper. A final section is also provided in which our experiments are extended and applied to two new datasets [20,35] for the sake of achieving generalisation.

2 Related Work

Several studies address the concept of the credibility of a webpage. Different teams have broadly analysed how online content credibility is assessed [10,26,40], and they have concluded that subjective ratings are very likely to rely on the user's background [26], e.g. their trust in technology, or on their reading skills [14].

Other researches focused on determining how the search engine result page (SERP) listings are used to determine credibility through user studies [22]. More specifically, several studies have been conducted related to assessing the credibility of health-related content on the web. For instance, Matthews et al. [25] analysed a corpus about alternative cancer treatments and found that almost 90% contained false claims. Liao and Fu [23] analysed age differences in credibility judgements and argued that older adults care less about the content of the site in comparison with younger ones.

Other teams focused on the association between different features and reliability. For example, Griffiths et al. [12] showed that algorithms like PageRank were unable to determine reliability on their own.

As can be seen, there are several concepts intimately related such as *reliabilty* [37], *trustworthiness* [20], *credibility* [35], or *veracity* [39]. Our reference study

Table 1. Class distribution in Sondhi's dataset.

	Sondhi et al.
# Reliable	180
% Reliable	50%
# Unreliable	180
% Unreliable	50%

will be Sondhi et al.'s [37] (which we will refer to from now on as the original paper), so we will use the same notion of reliability as them. For determining reliability, they defined their guidelines using the eight HONcode Principles[1]. For the generalisation experiments, we will consider the rest of the concepts (credibility, trustworthiness, etc.) as proxies of reliability (see Sect. 6).

3 Dataset

The original authors manually created a **fully balanced** dataset with reliable and unreliable webpages (see Table 1) that we directly used in our replicability task. This eases the classification task, but it is not very realistic since in real-world problems it is rare to find the same ratio among classes.

In the original paper, the authors randomly selected the positive pages from those websites accredited by HON[2] according to their principles. On the other hand, as HON does not report non-accredited sites, they searched the Web with a deliberate strategy to find poor quality pages. Using hand-crafted queries, such as *disease name + "miracle cure"*. To ensure that topical overlap between negative and positive instances (i.e. to avoid topic-bias classification), they conducted a topic analysis over the reliable corpus and extracted keywords related to diseases that occur in the set of reliable pages. For each keyword, they manually produced queries which involved terms like *treatment* or *miracle*. Finally, the authors checked and selected 180 unreliable pages from the search results. As the original download link for the dataset was no longer valid, the dataset was sourced via personal communication with the authors.

The main goal of the original paper was to build a **document-level classifier** using a standard supervised learning approach. We followed their experimental setup, in which the original authors argued that reliability can be represented as a binary value as the first approach to this problem.

3.1 Features

A variety of **features** were proposed based on style, content and external information such as links. As will be seen, we were not able to apply all of these in

[1] https://www.hon.ch/cgi-bin/HONcode/principles.pl?English.
[2] https://www.hon.ch/en/.

our experiments, since some tools or libraries were outdated, and other elements were not described in a sufficiently detailed way.

In the original paper, webpages were represented using several features, namely:

- **Link-based features:** the number and type of links are usually a good indicator of the type of website we are dealing with [4,5]. For example, as Sondhi and his colleagues exposed, a more reliable site tends to have more internal links, while a less reliable site tends to have more external links and advertisements [41]. On the other hand, the presence or absence of privacy policy information or contact links for the page author can be indicators of reliability. This is because the presence of these types of elements gives a sense of confidence to the user who consults the resource [11,21].
  Based on these criteria five features were defined to be taken into account: normalised value of internal links, normalised value of external links, normalised value of total links, the presence or not of contact link (boolean), and the presence or not of privacy link (boolean). For the latter two, the original paper did not explain how they were computed. Therefore, we manually defined two lists of privacy[3] and contact[4] expressions, such as *Privacy Policy* or *Contact Us*, after performing a first exploratory analysis over the documents.
  For normalisation, the original authors analysed a random sample of documents and they experimentally chose a large normalisation denominator (the link count was divided by Z_1, which was set to 200).
  In our experiments, the links were extracted from the text using the Beautiful Soup[5] Python package.
- **Commercial features:** the presence of commercial interest and advertisings often indicates a low reputation [4,41]. Therefore, two characteristics were defined to be taken into account: the normalised value of commercial links and the normalised frequency of commercial words on the website.
  For the latter, an initial list of indicative words of commercial interest was proposed in the article. We manually extend this list[6]. Since the original article was not explicit about word preprocessing, we followed a naive approach in which a word must match exactly with some of the words in the list to be taken into account in the final metric. This strategy can be improved in future versions by applying lemmatisation techniques, for example.
  Regarding normalisation, the normalised value of commercial links was obtained dividing by the same Z_1 used above. The second feature consisted of dividing the number of commercial words found by the document length.
- **PageRank Features:** the authors of the original paper used this feature as an indicator of the relative importance of a website [3]. However, this service has been removed by Google, and all Python packages that used their

[3] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/privacy.txt.
[4] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/contact.txt.
[5] https://www.crummy.com/software/BeautifulSoup/bs4/doc/.
[6] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/comm_list.txt.

endpoint cannot be applied. It would be still possible to manually compute PageRank based on the web graph. However, the current web graph does not reflect the situation of these pages when the collection was created (some pages are no longer accessible). Furthermore, previous work has shown that such features capture the popularity of a website, but fail to measure reliability [32].

- **Presentation features:** reliable content is usually presented carefully and clearly [11]. To evaluate this, the original paper employed *elinks*[7], a tool to extract the text of the webpage. Then, they defined two features based on the number of blank lines. However, in the final comparison, they did not include this feature set, so we did not take it into account in our replicability experiments.
- **Word-based features:** textual content and style are often good indicators of the reliability or reputation of a website [24,28]. Therefore, each word in a document was considered as a different dimension, taking its normalised frequency score. Since the original authors did not declare the use of any preprocessing stage, we applied no stemming or lemmatisation.
  We additionally considered two alternative pre-processing strategies, with and without *stopword* removal. To achieve this, the NLTK[8] English *stoplist* was manually extended[9] after a preliminary exploration of the documents.
  Finally, for each word we divided the number of occurrences of the word by the document length.

In addition to testing the feature sets in isolation, Sondhi and his colleagues also considered a final combination that merged **all features together**. In our case, we tested two variants of "all features" (one with word features extracted with stopword removal and another one with word features extracted with no stopword removal).

4 Experimental Setup

When carrying out the experimentation, a **vector support machine** was used as learning method. The original paper used a C++ implementation but, for compatibility reasons, we employed the SVMlight[10] Python wrapper. We are therefore facing a two-class classification problem.

To evaluate the results, we applied **5-fold cross validation**, as in the original study. When generating the predictions, there could be **two types of errors**: classifying a reliable page as non-reliable (FP) and classifying a non-reliable page as reliable (FN). The latter being the one we wish to avoid most. To make results comparable, the performance metric used is the same as in the original paper:

[7] http://elinks.or.cz.
[8] https://www.nltk.org/nltk_data.
[9] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/stopwords.txt.
[10] https://bitbucket.org/wcauchois/pysvmlight.

$$Weighted\ Accuracy(\lambda) = \frac{(\lambda \times TP) + TN}{\lambda \times (TP + FN) + TN + FP} \quad (1)$$

Three variants were considered, corresponding to $\lambda \in \{1, 2, 3\}$. Moreover, following the original paper strategy, the SVM classifier was trained with a cost-factor set to the value of λ (the weighted accuracy $\lambda = 1$ was obtained with a SVM whose cost-factor was set to 1, the weighted accuracy $\lambda = 2$ was obtained with a SVM whose cost-factor was set to 2, and so forth). Such an approach tunes the classifier to the measure that would later evaluate its effectiveness.

We note that the experiments were performed on an Ubuntu 19.04 machine, with 32 GB of RAM, 240 GB of storage and an Intel(R) Core(TM) i7-9750H CPU @ 2.60 GHz. The Python version used was 3.7.3 in an Anaconda 4.8.0 environment. However, for the CLEF eHealth dataset experiments, detailed in Sect. 6.2.2, it was necessary to use a server due to the storage requirements. More specifically, we used a CentOS 7.6.1810 machine, with 377 GB of RAM, 15T of storage and Intel(R) Xeon(R) CPU E5-2630 v4 processor. The Python and Anaconda versions used were the same as in the local experiments.

5 Results

Sondhi et al.'s original results are shown in Table 2. In our experiments, we considered two variants for word-based representation: with and without *stop-word* removal. Moreover, commercial features were not tested in isolation, but combined with link-based features. This is reasonable since they are intimately related to external and advertising links.

Our results (see Table 3) differ from the original ones, but the same conclusions can be drawn: word-based features and the merging all features achieve the best performance. Our comparison of the two word-based variants (with and without *stopwords*) suggests that keeping *stopwords* is the safest approach to estimate the reliability of a webpage.

We note that our best performance is higher than that obtained in the original work. More specifically, in our case, we observed a high increase in the performance obtained by merging all features together. This contrasts with the

Table 2. Sondhi et al. original paper results.

	Weighted accuracy (%)		
Features	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	60.8	71.1	79.6
Links + Commercial	67.8	75.9	79.6
Words	**80.6**	**83.9**	85.0
All	80.0	83.2	**86.8**

Table 3. Our results for Sondhi et al. dataset.

	Weighted accuracy (%)		
Features	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	70.5	80.0	73.5
Links + Commercial	69.7	79.4	74.3
Words (removing stopwords)	80.8	80.2	80.3
Words (keeping stopwords)	82.8	85.6	88.5
All (removing stopwords)	**97.5**	**98.3**	**98.6**
All (keeping stopwords)	96.1	96.3	96.5

original study, where the combination of features did not add value. This is perhaps the most surprising outcome of the replicability experiments, and the only plausible explanation we can derive is that this results from the setup differences between our experiments and the originals, as described in the previous sections.

6 Generalisation

To build on Sondhi et al.'s work and to determine the generalisability of their findings, we apply new **standardisation** techniques to the Sondhi et al. dataset and also test the methods with two **further datasets**.

6.1 Standardisation

The original paper authors did not report on how the **standardisation** of the features (to get 0 mean and 1 standard deviation) - commonly applied in machine learning [16] - could affect the algorithm performance. As such, we tested and report the results here (see Table 4).

As can be seen, the performance of all feature sets increases in comparison with results reported in Table 3. Of particular note, the models with word-based representation are most improved. By carrying out this procedure, in addition to the Z_1 normalisation per document previously described, we are favouring features or words that have a low average, that is, less-common or technical words (see Fig. 1). This evens out the differences between terms, and what really guides the classifier, is whether a feature of them deviates from its average in a particular document. For example, a word that is broadly used. This also explains why the best feature combination is word-based with *stopwords* being used.

6.2 New Test Datasets

The Web Search dataset by Schwarz et al. [35] and the CLEF eHealth consumer health search task 2018 [20] were used to further evaluate this classification technology. Both contain health-related content, but the first additionally addresses topics such as finance, politics, environment, and news about famous people.

Table 4. Our results for Sondhi et al. dataset (with standard scaler).

	Weighted accuracy (%)		
Features	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	74.4	78.1	76.4
Links + Commercial	73.3	76.5	79.9
Words (removing stopwords)	97.2	**98.3**	98.5
Words (keeping stopwords)	**98.1**	**98.3**	**98.9**
All (removing stopwords)	97.2	**98.3**	98.5
All (keeping stopwords)	97.8	**98.3**	**98.9**

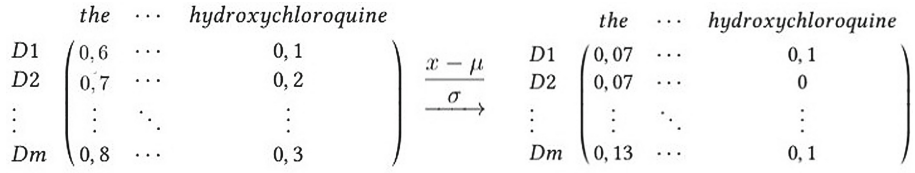

Fig. 1. Document-term matrix standardisation.

Schwarz et al. focused on credibility assessment to help people searching for information online. The CLEF eHealth task addresses a similar problem, but it is tighter to health-related online data. It must be noticed that these documents were not labelled in terms of reliability, but the notions of credibility and trustworthiness were used instead. However, we considered these concepts as proxies of reliability and attempted to see how generalisable the previous conclusions were against other datasets.

Schwarz et al. chose 1000 webpages related to multiple topics to be labelled in terms of credibility. They proposed a five-point Likert scale, from 1 to 5, to generate the ground-truth, and one of the authors of the paper rated the whole collection.

On the other hand, the CLEF eHealth consumer health search task dataset was created from webpages recovered from CommonCrawl[11]. The organisers of the task defined an initial list of potentially interesting sites and then, they submitted queries against a search engine to retrieve the final URLs. The initial list was extended by manually adding some reliable sites and other known to be unreliable. Finally, the corpus was divided into folders by domain.

In this CLEF task, it was decided to implement the RBP-based method proposed by Moffat et al. [27] to generate the assessment pool, instead of using a fixed-depth pooling strategy. After the pool was formed, human assessors from Amazon Mechanical Turk, with certain profiles, were selected. In the case of trustworthiness judgements, an eleven point scale, from 0 to 10, was used.

[11] http://commoncrawl.org.

It was necessary to relabel both datasets into a binary-class scale to fit with our 2-class technology. We removed the middle values (3 for Schwarz et al. and from 4 to 6 for CLEF) and mapped the extreme values to reliable and unreliable, respectively.

The main statistics of these datasets after performing this relabelling process are shown in Table 5. In both cases, we face an **imbalanced data** problem. This is particularly acute in the case of the Schwarz et al. data.

Table 5. Class distribution in the different datasets.

	Schwarz et al.	CLEF eHEALTH
# Reliable	75	9,879
% Reliable	93.75%	73.25%
# Unreliable	5	3,607
% Unreliable	6.25%	26.75%

Imbalanced learning is a common problem and there are multiple techniques to deal with the issue. In this case, we considered and compared two different approaches: introducing a **cost-factor** that applies a higher penalty to errors in the minority class and **resampling techniques** that try to balance the data by adding artificial instances or by removing some majority examples [6,15,17,18]. In this paper, only cost-factor techniques are reported since our preliminary experiments suggested that cost-factor methods outperform resampling methods in both datasets.

On the other hand, in imbalanced learning, it is common to use metrics, such as the **F1 measure**. Here, we report the micro-averaged F1, biased by the frequency of each class, and the value of F1 for each class. At the time of selecting the best feature combination for each collection, we gave priority to the minority class or unreliable F1.

Finally, it is worth noting that for both datasets the standardisation method described in Sect. 6.1 was applied.

6.2.1 Schwarz et al. Results

Due to the small dataset size, a stratified **2-fold cross validation** was used (instead of 5-folds). The obtained results are shown in Table 6. We note that in case of a tie, we always select the simplest feature set.

With **cost factor set to 1**, link-based features perform the best, but the classifier does not detect a single unreliable document. With this learning strategy, no combination is capable of correctly cataloguing examples from the minority class. This is not surprising given the low percentage of negative examples (6.25%).

Table 6. Our results for Schwarz et al. dataset.

					Weighted accuracy (%)		
Features	SVM cost factor	F1	F1 (reliable class)	F1 (non reliable class)	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	1	**0.94**	**0.97**	**0**	**93.75**	–	–
	2	0.94	0.97	0	–	88.26	–
	3	0.94	0.97	0	–	–	83.4
Links + Commercial	1	**0.94**	**0.97**	**0**	**93.75**	–	–
	2	0.94	0.97	0	–	88.26	–
	3	0.94	0.97	0	–	–	83.4
Words (removing stopwords)	1	0.93	0.96	0	92.5	–	–
	2	0.91	0.95	0.25	–	87.01	–
	3	**0.91**	**0.95**	**0.33**	–	–	**85.42**
Words (keeping stopwords)	1	0.91	0.95	0	91.25	–	–
	2	0.91	0.95	0	–	85.88	–
	3	0.91	0.95	0.2	–	–	84.54
All (removing stopwords)	1	**0.94**	**0.97**	**0**	**93.75**	–	–
	2	0.91	0.95	0	–	85.88	–
	3	0.91	0.95	0	–	–	81.13
All (keeping stopwords)	1	0.93	0.96	0	92.5	–	–
	2	**0.91**	**0.95**	**0.25**	–	**87.02**	–
	3	**0.91**	**0.95**	**0.33**	–	–	**85.42**

With **cost factor 2**, the results were still even, but some feature combinations were able to detect the minority class. This was the case of the word-based model and for the model combining all features- keeping *stopwords*. The latter was selected as the best combination, due to a slight difference in the weighted accuracy performance.

With **cost factor 3**, the detection of the minority class is slightly improved. As for the combination of features, both the word-based and the combination of all features (maintaining the *stopwords*) offer the same performance, but the former was selected because it generates a simpler model.

6.2.2 CLEF eHealth Results

This was the largest dataset in our experiments, and it also presents an imbalance problem between classes. In contrast with Schwarz et al., a stratified **5-fold cross validation** could be applied given the larger number of data points. The obtained results are shown in Table 7.

For all cost factor values, the word-based model that maintains the *stopwords* was the one that offered the best results, with also reasonable minority or non-reliable class detection.

6.3 Generalisation Conclusions

Each of the studied datasets was different both in terms of content and task. Moreover, the original collection was fully balanced, while the others were clearly imbalanced. Nevertheless, some interesting conclusions can be drawn from the generalisation experiments.

Table 7. Our results for CLEF eHealth dataset.

					Weighted accuracy (%)		
Features	SVM cost factor	F1	F1 (reliable class)	F1 (non reliable class)	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	1	0.73	0.85	0	73.15	–	–
	2	0.73	0.85	0	–	57.66	–
	3	0.46	0.39	0.28	–	–	50.39
Links + Commercial	1	0.73	0.85	0	73.15	–	–
	2	0.73	0.84	0	–	57.63	–
	3	0.3	0.12	0.41	–	–	51.74
Words (removing stopwords)	1	0.74	0.85	0.14	73.86	–	–
	2	0.68	0.79	0.38	–	61.57	–
	3	0.55	0.63	0.44	–	–	58.65
Words (keeping stopwords)	1	**0.75**	**0.85**	**0.24**	**74.63**	–	–
	2	**0.69**	**0.79**	**0.41**	–	**62.93**	–
	3	**0.59**	**0.68**	**0.45**	–	–	**59.81**
All (removing stopwords)	1	0.74	0.85	0.15	73.88	–	–
	2	0.68	0.79	0.38	–	61.58	–
	3	0.55	0.62	0.44	–	–	58.39
All (keeping stopwords)	1	0.75	0.85	0.24	74.53	–	–
	2	0.7	0.79	0.4	–	62.89	–
	3	0.59	0.67	0.45	–	–	59.72

The obtained results **reinforce** the main insights of the original study. In all of the experiments the best strategies are the bag-of-words approach or the one that merges all features set together. The evidence moreover suggests that keeping stopwords leads to enhanced performance.

7 Future Work

This work opens up a line of research that allows us to continue to study in-depth how unreliable information is transmitted in the Web and how it is perceived by users. A natural next step would be the application of our predictive technology to the case of **social media** [1,13,38], extracting known true and false claims from the labelled documents and seeing their impact on this media. This kind

of news spreads very quickly in this media, which can help us to identify them or put them under suspicion.

We also intend to further analyse the effect of combining different features on performance and, additionally, plan to train new models using **BERT** [7]. This language modelling approach, which extracts a contextual representation of words, has been proven to be successful in the field of Natural Language Processing (NLP).

We will also perform transfer learning experiments among the different datasets available [8,29]. This can be helpful to understand whether or not training with one collection and testing with another reinforces the conclusions obtained.

8 Conclusions

In this work, a replicability study of reliability technology was presented. The main objective was to re-run the experiments and try to confirm the conclusions extracted from the original study. Our results reinforce the fact that word-based models or the ones that combine all available features are the most promising approaches to distinguish reliable from unreliable sites.

We have also tested this predictive technology against two further and highly different datasets and the conclusions remain the same. This gives us the confidence to state that the research presented in the original paper establishes a good reference for reliability detection in online data.

Finally, as a new test of its generalisation, this algorithm has been used by our team in the TREC 2020 Health Misinformation Track[12] to tackle misinformation about COVID-19 and its treatments. In order to replicate the experiments presented in this work, the code is available for the research community at Github[13].

Acknowledgements. This work was funded by FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación/Project (RTI2018-093336-B-C21). This work has received financial support from the Consellería de Educación, Universidade e Formación Profesional (accreditation 2019–2022 ED431G-2019/04, ED431C 2018/29, ED431C 2018/19) and the European Regional Development Fund (ERDF), which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System.

References

1. Abbasi, M.-A., Liu, H.: Measuring user credibility in social media. In: Greenberg, A.M., Kennedy, W.G., Bos, N.D. (eds.) SBP 2013. LNCS, vol. 7812, pp. 441–448. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37210-0_48

[12] https://trec-health-misinfo.github.io.
[13] https://github.com/MarcosFP97/Health-Rel.

2. Abualsaud, M., Smucker, M.D.: Exposure and order effects of misinformation on health search decisions. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Rome (2019)
3. Andersen, R., et al.: Robust pagerank and locally computable spam detection features. In: Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, pp. 69–76 (2008)
4. Becchetti, L., Castillo, C., Donato, D., Baeza-Yates, R., Leonardi, S.: Link analysis for web spam detection. ACM Trans. Web (TWEB) **2**(1), 1–42 (2008)
5. Borodin, A., Roberts, G.O., Rosenthal, J.S., Tsaparas, P.: Link analysis ranking: algorithms, theory, and experiments. ACM Trans. Internet Technol. (TOIT) **5**(1), 231–297 (2005)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Do, C.B., Ng, A.Y.: Transfer learning for text classification. Adv. Neural Inf. Process. Syst. **18**, 299–306 (2005)
9. Eysenbach, G.: Infodemiology: the epidemiology of (mis)information. Am. J. Med. **113**(9), 763–765 (2002)
10. Fogg, B.J.: Prominence-interpretation theory: explaining how people assess credibility online. In: CHI 2003 Extended Abstracts on Human Factors in Computing Systems, pp. 722–723 (2003)
11. Ginsca, A.L., Popescu, A., Lupu, M.: Credibility in information retrieval. Found. Trends Inf. Retr. **9**(5), 355–475 (2015). https://doi.org/10.1561/1500000046
12. Griffiths, K.M., Tang, T.T., Hawking, D., Christensen, H.: Automated assessment of the quality of depression websites. J. Med. Internet Res. **7**(5), e59 (2005)
13. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: TweetCred: real-time credibility assessment of content on Twitter. In: Aiello, L.M., McFarland, D. (eds.) SocInfo 2014. LNCS, vol. 8851, pp. 228–243. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13734-6_16
14. Hahnel, C., Goldhammer, F., Kröhne, U., Naumann, J.: The role of reading skills in the evaluation of online information gathered from search engine environments. Comput. Hum. Behav. **78**, 223–234 (2018)
15. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: review of methods and applications. Expert Syst. Appl. **73**, 220–239 (2017)
16. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. SSS. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
17. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
18. Hoens, T.R., Chawla, N.V.: Imbalanced datasets: from sampling to classifiers. Imbalanced Learning: Foundations, Algorithms, and Applications, pp. 43–59 (2013)
19. Islam, M.S., et al.: Covid-19-related infodemic and its impact on public health: a global social media analysis. Am. J. Trop. Med. Hyg. **103**(4), 1621–1629 (2020)
20. Jimmy, J., Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L.: Overview of the CLEF 2018 consumer health search task. In: International Conference of the Cross-Language Evaluation Forum for European Languages (2018)

21. Kakol, M., Nielek, R., Wierzbicki, A.: Understanding and predicting web content credibility using the content credibility corpus. Inf. Process. Manag. **53**(5), 1043–1061 (2017)
22. Kattenbeck, M., Elsweiler, D.: Understanding credibility judgements for web search snippets. Aslib J. Inf. Manag. **71**, 368–391 (2019)
23. Liao, Q.V., Fu, W.T.: Age differences in credibility judgments of online health information. ACM Trans. Comput.-Hum. Interact. (TOCHI) **21**(1), 1–23 (2014)
24. Matsumoto, D., Hwang, H.C., Sandoval, V.A.: Cross-language applicability of linguistic features associated with veracity and deception. J. Police Crim. Psychol. **30**(4), 229–241 (2015)
25. Matthews, S.C., Camacho, A., Mills, P.J., Dimsdale, J.E.: The Internet for medical information about cancer: help or hindrance? Psychosomatics **44**(2), 100–103 (2003)
26. McKnight, D.H., Kacmar, C.J.: Factors and effects of information credibility. In: Proceedings of the Ninth International Conference on Electronic Commerce, pp. 423–432 (2007)
27. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. (TOIS) **27**(1), 1–27 (2008)
28. Mukherjee, S., Weikum, G.: Leveraging joint interactions for credibility analysis in news communities. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 353–362 (2015)
29. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2009)
30. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J.G., Rand, D.G.: Fighting Covid-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. Psychol. Sci. **31**(7), 770–780 (2020)
31. Pogacar, F.A., Ghenai, A., Smucker, M.D., Clarke, C.L.: The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, pp. 209–216 (2017)
32. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 2173–2178 (2016)
33. Reuters Institute, University of Oxford: Reuters Digital News Report 2020 (2020). https://www.digitalnewsreport.org/survey/2020. Accessed 16 Nov 2020
34. Rieh, S.Y.: Judgment of information quality and cognitive authority in the web. J. Am. Soc. Inf. Sci. Technol. **53**(2), 145–161 (2002)
35. Schwarz, J., Morris, M.: Augmenting web pages and search results to support credibility assessment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1245–1254 (2011)
36. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y.: Combating fake news: a survey on identification and mitigation techniques. ACM Trans. Intell. Syst. Technol. (TIST) **10**(3), 1–42 (2019)
37. Sondhi, P., Vydiswaran, V.G.V., Zhai, C.X.: Reliability prediction of webpages in the medical domain. In: Baeza-Yates, R., et al. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 219–231. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28997-2_19
38. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health information-a survey. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. **7**(5), e1209 (2017)

39. Vydiswaran, V.V., Zhai, C., Roth, D.: Content-driven trust propagation framework. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 974–982 (2011)
40. Yamamoto, Y., Tanaka, K.: Enhancing credibility judgment of web search results. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1235–1244 (2011)
41. Zha, W., Wu, H.D.: The impact of online disruptive ads on users' comprehension, evaluation of site credibility, and sentiment of intrusiveness. Am. Commun. J. **16**(2), 15–28 (2014)