Health Misinformation Detection in Web Content

A Structural-, Content-based, and Context-aware Approach based on Web2Vec

Rishabh Upadhyay
r.upadhyay@campus.unimib.it
University of Milano-Bicocca
Milan, Italy

Gabriella Pasi
gabriella.pasi@unimib.it
University of Milano-Bicocca
Milan, Italy

Marco Viviani
marco.viviani@unimib.it
University of Milano-Bicocca
Milan, Italy

ABSTRACT

In recent years, we have witnessed the proliferation of large amounts of online content generated directly by users with virtually no form of external control, leading to the possible spread of misinformation. The search for effective solutions to this problem is still ongoing, and covers different areas of application, from opinion spam to fake news detection. A more recently investigated scenario, despite the serious risks that incurring disinformation could entail, is that of the online dissemination of health information.

Early approaches in this area focused primarily on user-based studies applied to Web page content. More recently, automated approaches have been developed for both Web pages and social media content, particularly with the advent of the COVID-19 pandemic. These approaches are primarily based on handcrafted features extracted from online content in association with Machine Learning. In this scenario, we focus on Web page content, where there is still room for research to study structural-, content- and context-based features to assess the credibility of Web pages.

Therefore, this work aims to study the effectiveness of such features in association with a deep learning model, starting from an embedded representation of Web pages that has been recently proposed in the context of phishing Web page detection, i.e., Web2Vec.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; • **Information systems** → **Web mining**; **Document representation**.

KEYWORDS

Health Misinformation, Credibility, Social Web, Machine Learning, Deep Learning.

ACM Reference Format:
Rishabh Upadhyay, Gabriella Pasi, and Marco Viviani. 2021. Health Misinformation Detection in Web Content: A Structural-, Content-based, and Context-aware Approach based on Web2Vec. In *Conference on Information Technology for Social Good (GoodIT '21), September 9–11, 2021, Roma, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3462203.3475898

GoodIT '21, September 9–11, 2021, Roma, Italy

ACM ISBN 978-1-4503-8478-0/21/09…$15.00
https://doi.org/10.1145/3462203.3475898

1 INTRODUCTION

The problem of widespread online misinformation has given impetus in recent years to the proliferation of research works that have attempted to curb this phenomenon from different perspectives and with respect to different domains. Distinct approaches have been applied mainly to review sites, i.e., online platforms that allow users to publish reviews about products and services [28], and microblogging platforms, which often disseminate newsworthy content related to politics and events [44]. The majority of these solutions are often based on the identification of particular characteristics (i.e., *features*) that are highly domain-specific, related to the content being disseminated, the purpose of the dissemination, the platform being considered, the authors of the content, and possible social interactions in the case of social networking sites. Such features are often considered within supervised classifiers categorizing genuine versus non-genuine information, using "standard" Machine Learning or more recent deep learning models [16, 40].

However, one domain that has been less considered in developing automated solutions for evaluating information credibility, is that of online health-related content dissemination, despite the severe harm one might incur in coming into contact with misinformation when searching for possible health treatments and advice. In [8], *health misinformation* has been defined as: "a health-related claim of fact that is currently false due to a lack of scientific evidence". In most cases, people who are not an expert in the field are unable to properly assess the reliability of such claims, both, in general, due to their limited cognitive abilities [27] and, more specifically, due to their insufficient level of *health literacy* [38]. The difficulties in providing people with automated solutions to compensate for the complexity of evaluating health information on their own, in an online context that is less and less mediated by the presence of medical experts [12], lie in the fact that online content related to the health domain has its own peculiarities compared to other domains of interest. First of all, both long and semi-structured texts published on "traditional" Web pages (e.g., forums, blogs, question-answering medical systems, etc.), and very short and unstructured texts spread through microblogging platforms (e.g., the mass of COVID-19-related tweets in the last year) are diffused online. Secondly, these texts are characterized by a scientific language and possible reference to external resources that can be taken into account when assessing their credibility.

With the aim of contributing to social good in the context of studying solutions to prevent people from coming into contact with potentially harmful health misinformation, this work focus on health-related content disseminated in the form of Web pages, an area in which research has mainly identified some (handcrafted) features that make a site or a page "credible", through the use of

user-based studies or ML approaches. In this article, the possibility of representing Web pages by means of automatically learned embedding features is explored by considering Web2Vec, a solution recently proposed for phishing Web page detection [14]. With respect to Web2Vec, we inject some credibility aspects in the feature extraction phase and in the deep learning architecture employed. Evaluations are performed against publicly available datasets containing health-related content in the form of Web pages, and some baselines that have been proposed in the literature to assess the credibility of the information in both the general and health fields.

2 RELATED WORK

Given the purpose of this article, in this section we focus primarily on describing approaches that have considered the problem of assessing the credibility of health-related Web pages, while mentioning, however, some recent research directions with respect to social platforms.

2.1 Interactive-based Approaches

Initially, so-called *interactive-based* approaches applied to Web sites/pages have been considered to assess the credibility (and/or other related concepts partially overlapping or tangent to that of credibility) of health-related information, by employing interview questionnaires or other interactions with the users. Such approaches are based on users' *subjective* perceptions, which are driven by their information needs and other personal/demographic factors [11, 37]. In [20, 32], the authors stated that users' evaluations of Web content are influenced, in particular, by *source-related*, *content-related*, and *design-related factors*.

Such factors can have either positive or negative effects on the users' evaluations with respect to the credibility of information. For example, source-related factors such as the authority of owners/sponsors have been shown to have a positive effect on the perception of credibility [24], even if, in some cases, users tend not to trust communication that is too "institutional" [41], perceived in particular by younger users as old and "not cool" [29]. Whereas for content-related factors, the consensus among sources has been considered as a good credibility indicator. Users have shown mixed feeling towards personal experience and facts. Some users evaluate the presence of "objective" facts in a positive way [41], while others find imbalance in their presence [33]. Design-related factors can be evaluated either positively or negatively: in general, an aesthetically well-maintained and easy-to-navigate site is perceived as credible. On the contrary, it is perceived negatively [26].

However, the credibility of online health-related information is a complex concept involving more than two dozen dimensions subjectively assessed by users. For this reason, in recent years, some approaches have been proposed to automatically address the problem both with respect to Web and social media content.

2.2 Automated Approaches

In [43], the authors presented a framework for predicting the so-called *resource quality* (RQ) of medical Web pages, by using an SVM model trained on 750 resources published on Breast Cancer Knowledge Online (BCKOnline) [23]. The authors described RQ as a composition of *reliability* and *relevance*, where reliability is assessed based on quality dimensions such as *accuracy*, *credibility*, and *currency* [42], and relevance is related to the utility of the page for a user searching for a given medical information. The considered features are *categorical features*, i.e., related to the audience (e.g., age, disease stage, etc.), the type (e.g., medical, supportive, and personal), and the subject of the Web page (e.g. treatment-related, therapies-related, etc.), and simple *textual features*, such as title, description, creator, publisher, and access right.

In [36], an automatic approach based on SVM for *reliability* prediction of medical Web pages has been proposed. In the approach, the reliability of a Web page is assessed by performing binary classification. The author explored the usage of *link-based*, *commercial*, *PageRank*, *presentation*, and *textual features*. Link-based features are related to some counting of internal and external links (and related properties) in the Web page. Commercial features refer to the presence of commercial terms in the page. PageRank features are related to the relative importance of a Web page computed via PageRank.[1] Presentation features refer to the clearly in the presentation of content on the page. Finally, textual features are simply defined as normalised word frequency vectors.

Two other recent works based on the use of handcrafted features and Machine Learning approaches are those described in [15, 25]. In [25], a Logistic Regression model for assessing the *reliability* of Web pages has been trained on labeled data collected w.r.t. 13 vaccine-related search queries. *Textual features* are employed in the form of count-based and TF-IDF word vectors. In [15], a replicability study has been conducted on [36], considering two additional datasets made available in [34, 39], and ignoring PageRank features, deemed as not suitable for assessing Web content reliability [30].

Solutions that attempt to refer to criteria of reliability of medical information provided by external bodies are those proposed by [1, 5, 9, 21]. In [5], the capability of an automated system to perform the task of identifying 8 HONcode principles on health Web sites has been studied.[2] Distinct Naive Bayes classifiers are trained over a collection of Web pages labeled w.r.t. the considered criteria. In this approach, Web site content is converted into weighted bag-of-words representations. In [1], to confirm the *evidence-based medicine* (EBM) property of a Web page, the treatment described in the Web page is checked w.r.t. its approval by the US Food and Drug Administration, the UK National Health System, or the National Institute of Care Excellence. Two different feature types are considered for classification: *text-based* and *domain-specific* features, related to JAMA criteria [35]. A number of distinct ML classifiers have been used for the experiments. In [21], the authors have proposed to automate the use of DISCERN.[3] Five *Hierarchical Encoder Attention-based* (HEA) models (related to 5 DISCERN criteria) are trained on articles related to breast cancer, arthritis, and depression. A *Bidirectional Recurrent Neural Network* (BRNN) layer converts words, sentences, and documents to dense vector representations. Such representations are used for classification using a softmax layer. A knowledge-guided graph attention network named DETERRENT for detecting *health misinformation* has been recently

[1] https://metacpan.org/release/WWW-Google-PageRank/
[2] The HONcode certification is an ethical standard aimed at offering quality health information. https://www.hon.ch/cgi-bin/HONcode/principles.pl?English
[3] DISCERN is a questionnaire providing users with a way of assessing the quality of information on health treatment choices. http://www.discern.org.uk/

proposed in [9], trained on articles related to diabetes and cancer. It incorporates a Medical Knowledge Graph and an Article-Entity Bipartite Graph, and propagates node embeddings representing Web pages through Knowledge Paths for misinformation classification.

More recently, the interest of the scientific community to detect health misinformation is turning to the use of *deep learning* solutions, especially w.r.t. social media content [3, 31], achieving promising results. However, w.r.t. the tangent problem of phishing Web page detection, a recent deep learning approach presented in [14], i.e., Web2Vec, has been proposed. Hence, in this paper we intend to study such a solution (suitably modified) with respect to the problem of assessing the credibility of Web page content, by considering, in the deep learning model, source, content, and design factors, together with contextual aspects related to the presence of links and medical-related terms in the Web page.

3 A WEB2VEC-BASED SOLUTION FOR HEALTH MISINFORMATION DETECTION

The original Web2Vec model [14], developed for phishing Web page detection, is based on the embedded representation of the URL, content, and DOM structure of the considered Web page. Such embedding representations are used by a hybrid CNN-BiLSTM network to extract local and global features, which are combined by an attention mechanism strengthening important features. Multi-channel output vectors are concatenated and provided to a classifier to determining the category of the tested Web page (i.e., phishing vs non-phishing).

In the approach proposed in this article, in the application and in the appropriate modification of the Web2Vec model, some characteristics related to the problem of assessing the credibility of health information are taken into account. First of all, when generating an embedded representation of Web pages, the use of a specific vocabulary related to the medical field is considered, which is crucial to detect health misinformation. In addition, instead of focusing on the features related to the URL of the Web page to be evaluated (as Web2Vec did), those related to the URLs present in the page itself are considered, because they can give a better indication of whether they refer to reliable or unreliable external sources (e.g., the presence of commercial links). With these aspects in mind, the proposed solution consists of the following phases:

- *Data Parsing*: page links, content, and Document Object Mode (DOM) structure are parsed from each HTML page in the dataset, to extract suitable data that are employed in the following phase;
- *Data Representation*: word-level and sentence-level embedding representations are generated for the Web page content, while for the DOM structure and links, HTML tags and URLs embeddings are considered;
- *Feature Extraction*: a CNN-BiLSTM network is used to extract features from the given representations;
- *Web Page Classification*: health-related Web pages are classified as credible or not credible by using densely connected layers.

3.1 Data Parsing

The data parsing operation is the same as in Web2Vec, with the exception of link parsing, which in the case of this approach is applied to the content of the HTML page.

DOM Corpus. HTML files are characterized by a typical semi-structured data format. The hierarchical structure is represented using HTML tags, organized according to the Document Object Model (DOM) structure. Focusing on such a structure, we extracted an ordered list of tags, starting from high-level tags until "children" tags, i.e., `HTML`, `HEAD`, `META`, `LINK`, `TITLE`, `SCRIPT`, `BODY`, `DIV`, `TABLE`, `TR`, `TD`, `IMG`. Such HTML tags are considered as words, which constitute the *world-level corpus* for the DOM structure to be used in the data representation phase.

Content Corpus. Each Web page is phrased, and only textual content is considered (links and tags are excluded). Both a *word-level* and a *sentence-level* corpus are constructed. The first is constituted by each distinct word present in the page, the second identifies word sequences separated by the ' . ' character. Specifically, we consider a fixed-length dimension for each word sequence.

Link Corpus. The link corpus is created considering links present in the HTML page. In particular, we focus on the domain names extracted from the URL of the Web sites referenced in the HTML page. Such domain names, illustrated in Figure 1, represent the *word-level corpus* to be employed in the data representation phase.

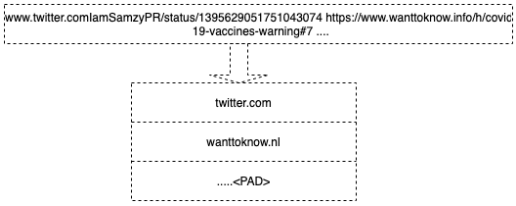

Figure 1: Construction of the word-level corpus for links.

3.2 Data Representation

In this phase, the word- and sentence-level corpora related to Web page DOM structure, content, and links, generated in the previous phase, are formally represented in order to capture their semantic relationships through word embedding. In particular, a Keras embedding layer is employed,[4] which is based on a supervised method that enhance the semantic representation while training the model using backpropogation. It is worth to be underlined that a separate embedding layer is defined for the DOM corpus, the word-level content and link corpora, and the sentence-level content corpus.

With respect to Web2Vec, in this work, to include domain-specific information related to the medical field, we add a word2vec layer pre-trained on PubMed as a weight initializer in the Keras embedding layer when considering the content word-level embedding. In this way, the word2vec weights are used as weight initializers for the embedding layer, as illustrated in Figure 2.

3.3 Feature Extraction

The features, as in the case of Web2Vec, are extracted by means of a CNN-BiLSTM network with an attention mechanism applied to the embedding representations obtained in the previous phase. *Convolutional Neural Networks* (CNN)s are nowadays commonly

[4] https://keras.io/api/layers/core_layers/embedding/

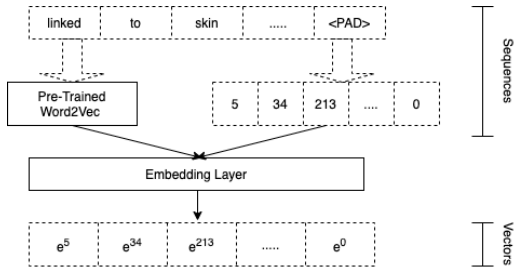

Figure 2: The word-level embedding phase for the content.

used for local feature extraction from data. *Bidirectional Long Short-Term Memory* (Bi-LSTM) networks are used to overcome the ability to learn feature from sequences of CNN by combining the word with its context [13]. The attention mechanism is used to improve the prediction capacity of the model.

CNN. The employed CNN is constituted, as in Web2Vec, by a feedforward network model structure. The hidden layer is divided into a convolution layer and a pooling layer. To overcome overfitting, each fully connected layer is followed by one dropout layer (with a dropout ratio of 0.05). Details on the convolution and pooling operations can be found in [14].

BiLSTM. The output of the CNN layer constitutes the input of the BiLSTM layer. Such layer is formed using Long Short-Term Memory in both directions, i.e., forward and backward, which keeps the sequential order among the data. It also allows detecting the relationship between the previous inputs and the output.

Since BiLSTM is a sequential- and memory-based model, it can both learn long-term dependence on the Web page and also extract improved features using local features from CNN. To deal with possible overfitting, *dropout learning* and *L2 regularization* (as detailed in the next section) are used to improve the model training.

Attention Layer. The addition of the attention layer, in the case of assessing the credibility of health-related information, is dictated by the fact that in the same document there may be parts characterized by "more credible" and "less credible" information. In this situation, even the presence of a small amount of "non-credible" features characterizing a credible page (or vice versa), can negatively affect its final evaluation. The purpose of the attention layer is, therefore, to pay particular attention with respect to the most discriminant features w.r.t. the considered problem; in this work, we have referred in particular to the concept of *additive attention* [2].

3.4 Web Page Classification

Web pages are categorized into credible and not credible through the use of a *binary classifier* consisting of a fully connected layer having a sigmoid function in the final layer, which combines the features extracted from the previous layers relating to the four corpora considered (DOM corpus, word-level content and link corpora, sentence-level content corpus).

For the classification loss calculation, the *cross-entropy loss function* and the *L2 regularization* are applied to overcome overfitting. Formally: $Error(t-y) = -\frac{1}{N}\sum_n^N [t_n \ln y_n + (1-t_n)\ln(1-y_n)]$, and $Loss = Error(t-y) + \lambda \sum_n^N w_n^2$, where t is the target label, y the predicted label, w the weight matrix of the layer, and λ is the so-called *L2 penalty parameter*.

4 EXPERIMENTAL EVALUATIONS

In this section, we present the experimental evaluation of the effectiveness of the proposed model. Specifically, we introduce the description of the different datasets, baselines, and evaluation metrics considered, together with technical and experimental details followed by a discussion on the obtained results.

4.1 Description of the Datasets

Only a few publicly available datasets are currently available for evaluating health-related information w.r.t. credibility. In particular, it has been necessary to consider those datasets from which it was possible to obtain the original HTML format of Web pages. Hence, the choice has fallen on the datasets provided by [18, 34, 36].

Microsoft Credibility Dataset [34]. This dataset is constituted by 1,000 Web pages in different domains such as Health, Finance, Politics, etc. Credibility ratings associated with them are provided over a five-point Likert scale, ranging from 1 to 5, where 1 stands for "very non-credible", and 5 for "very credible". In [15], for evaluation purposes, labels have been pre-processed by removing the middle value 3, and mapping 4-5 rating values to credible Web pages and 1-2 rating values to non-credible Web pages. In our approach, we followed the same strategy, and we focused on the 130 available health-related Web pages. Out of 130, 104 were credible and 26 were non-credible. Given the high data imbalance, we applied the SMOTE [7] oversampling method to the minority class.

Medical Web Reliability Corpus [36]. This is a manually generated balanced dataset with binary (i.e., reliable and unreliable) labels associated with Web pages. The authors randomly selected reliable Web sites from HON accredited Web sites.[5] Unreliable Web sites were searched on the Web using queries, constituted by the disease name + "miracle cure". The dataset consists of 360 Web pages, 180 reliable and 180 unreliable. After a cleaning phase, to remove blank and no-longer accessible pages, we dealt with 170 reliable Web pages and 176 unreliable Web pages.

CLEF eHealth 2020 Task-2 Dataset [18]. This dataset consists of a larger number of documents than the previously illustrated datasets, and has been expressly built to assess the topical relevance, readability, and credibility of Web pages consisting of medical content, as part of the so-called *Consumer Health Search* (CHS) task.[6] Credibility ratings are expressed on a four-point scale, from 0 to 3. Such ratings have been converted to binary values by considering 0-1 values as non-credible and 1-2 values as credible. Finally, we dealt with 5,509 credible and 6,736 non-credible Web pages.

4.2 Baselines and Evaluation Metrics

The approaches that are taken into consideration as baselines for assessing the effectiveness of the proposed approach concern solutions developed for assessing the credibility of both "general"

[5] https://www.hon.ch/en/
[6] https://clefehealth.imag.fr/?page_id=610

and health-related information, which consider both textual and other families of handcrafted features in association with Machine Learning. In particular, we consider the textual-feature-based model proposed in [25], the multi-feature-based model proposed in [15], which encompasses another multi-feature-based model discussed in [36], and a BioBERT-SVM model that has been developed in this work for evaluation purposes, given that BERT embeddings have produced a good result in association with SVM in fake news and misinformation detection problems [10, 17, 19]. Specifically, with respect to this baseline, we consider BERT embeddings pre-trained on PubMed articles for adapting to the biomedical domain [22].

With respect to the above-mentioned baselines, the following evaluation metrics are taken into consideration: F1 measure, accuracy and AUC. Such metrics have often been used in various literature works related to misinformation detection and credibility assessment [9, 25]. For training the ML models employed as baselines, the `scikit-learn` library [6] has been used.[7] To evaluate the results, 5-fold stratified cross-validation has been applied.

4.3 Results and Discussion

This section illustrates and discusses the results of the proposed solution with respect to each dataset and baseline described in the previous sections, in terms of the above-mentioned evaluation metrics respectively. In the following, the considered *baselines* are denoted as: NB-CountVec and LR-TF-IDF, identifying the most effective approaches presented in [25], based on the application of a Naive Bayes and a Logistic Regression classifier to textual features expressed as count vectors and TF-IDF vectors; MFB-SVM, denoting the multi-feature model presented in [15]; and BioBERT-SVM, as detailed in Section 4.2. Furthermore, two Web2Vec variations have been considered:[8]

- Web2Vec(C): it refers to the Web2Vec model trained only on content embeddings with default weight initializers;
- Web2Vec(C-D): it refers to the Web2Vec model trained on both content and DOM embeddings with default weight initializers.

Such additional baselines have been compared with distinct instantiations of the proposed model based on Web2Vec for assessing credibility, denoted as Cred-W2V. In particular:

- Cred-W2V(C): it refers to the proposed model trained on content embeddings with the PubMed word2vec layer acting at weight initializer;
- Cred-W2V(C-D): it refers to the proposed model trained on content embeddings with the PubMed word2vec layer, and on DOM embeddings with default weights;
- Cred-W2V(C-D-L): it refers to the proposed model trained on content, DOM, and link embeddings with default weight initializers;[9]
- Cred-W2V(C-D-L)*: it refers to the proposed model trained on content, DOM, and link embeddings, with the PubMed word2vec layer acting at weight initializer.

[7] https://scikit-learn.org/

[8] It was not possible to evaluate Web2Vec also w.r.t. its original link embedding model because in the considered datasets the URLs of the considered pages were missing.

[9] With respect to the original Web2Vec model, we recall that links are referred to those URLs present in a Web page.

The considered datasets (Section 4.1), are denoted as D1 (Microsoft Credibility Dataset), D2 (Medical Web Reliability Corpus), and D3 (CLEF eHealth 2020 Task-2 Dataset). Only for D3, it was possible to calculate, given the higher number of labeled data, the *binomial proportion confidence intervals* with 95% confidence [4]. In this case, the results are reported under the label D3(BI).

Table 1: Evaluation results.

	Metrics	D1	D2	D3	D3(BI)
NB-CountVec	Accuracy	74.55	94.43	64.89	64.9 ± 3.00
	F1	83.22	94.71	67.84	67.2 ± 3.00
	AUC	67.02	93.98	64.12	64.6 ± 2.93
LR-TF-IDF	Accuracy	75.35	94.29	68.6	67.9 ± 2.55
	F1	85.82	94.37	71.3	70.9 ± 2.80
	AUC	47.18	93.21	67.6	67.8 ± 2.55
MFB-SVM	Accuracy	70.03	94.73	66.15	63.8 ± 3.50
	F1	75.97	93.52	46.03	46.7 ± 2.50
	AUC	57.44	93.98	47.78	50.2 ± 0.10
BioBERT-SVM	Accuracy	72.1	94.1	70.74	69.8 ± 2.00
	F1	44.67	94.2	65.34	65.3 ± 4.00
	AUC	63.2	94.1	69.56	67.0 ± 3.00
Web2Vec(C)	Accuracy	*78.34*	*94.81*	70.34	69.5 ± 2.50
	F1	*85.67*	*94.49*	*71.56*	*68.9 ± 2.75*
	AUC	*65.34*	*94.54*	*70.18*	*68.9 ± 2.10*
Cred-W2V(C)	Accuracy	*78.34*	***96.1***	***71.38***	***71.5 ± 1.75***
	F1	***86.34***	***95.21***	***72.35***	***71.8 ± 2.25***
	AUC	***68.13***	***95.98***	***71.59***	***70.9 ± 2.10***
Web2Vec(C-D)	Accuracy	*80.7*	*96.4*	*72.12*	*71.9 ± 2.22*
	F1	*88.28*	*96.12*	*73.69*	*72.5 ± 1.70*
	AUC	*74.34*	*96.32*	*71.71*	*71.1 ± 1.75*
Cred-W2V(C-D)	Accuracy	***86.9***	***97.57***	***73.58***	***72.5 ± 2.20***
	F1	***91.62***	***97.69***	***77.98***	***75.5 ± 2.15***
	AUC	***80.07***	***97.42***	***73.59***	***72.8 ± 1.40***
Cred-W2V(C-D-L)	Accuracy	***84.12***	***96.23***	***73.98***	***73.4 ± 1.70***
	F1	***90.45***	***96.24***	***75.74***	***75.1 ± 1.97***
	AUC	***78.17***	***96.26***	***73.85***	***73.4 ± 2.10***
Cred-W2V(C-D-L)*	Accuracy	***90.75***	***98.81***	***74.42***	***75.5 ± 2.35***
	F1	***94.17***	***97.77***	***76.62***	***78.5 ± 2.15***
	AUC	***86.17***	***97.68***	***74.24***	***75.8 ± 1.50***

As it can be seen from Table 1, the proposed model for health misinformation detection outperforms all the baselines that rely on the use of handcrafted features and Machine Learning techniques with respect to all datasets and evaluation metrics considered (values in "italic"). Also compared to using the original Web2Vec model applied to the problem considered in this article, the proposed model allows to obtain better results (values in "bold"), both when the word2vec layer trained on PubMed is added to the original architecture, and when we consider the embeddings of the links present in the pages to be evaluated. We can say in particular, looking at the comparison of the results of the Cred-W2V(C-D), Cred-W2V(C-D-L) and Cred-W2V(C-D-L)* models, that the impact of including a pre-trained embedded representation on a domain-specific lexicon is preponderant over the effectiveness of the proposed approach. However, this may be due to the fact that, in the considered datasets, some HTML pages presented no internal links, in some cases due to the original gathering process itself.

5 CONCLUSIONS

In this article, we addressed the issue of identifying misinformation in health-related Web pages. Starting from an analysis of the literature works, we identified some of the most frequently adopted features and solutions to address the considered problem. In particular, "classical" Machine Learning techniques and a deep learning approach recently proposed for detecting phishing Web pages, i.e., Web2Vec. Starting from this model, we have applied and suitably modified it w.r.t. the problem of health misinformation detection, concluding that such an approach can be effective when customized on the considered domain and, in any case, it is more effective than traditional ML-based methods.

This work represents a first step for the investigation at a more general level of what can be actually the most effective architectures and features to solve the problem; from the work it emerged that actually a semantic and context-aware representation of the text contained in the Web pages has a big impact on the effectiveness of the identification of health misinformation. However, the distinct impact of the presence of links and structural features of the page is certainly worth investigating in the future, together with the possibility to insert into the model domain knowledge also through the addition of both handcrafted and other embedding features.

ACKNOWLEDGEMENTS

This work is supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval (H2020-EU.1.3.1., ID: 860721).

REFERENCES

[1] Majed M. Al-Jefri et al. Using machine learning for automatic identification of evidence-based health information on the Web. In *ACM Int. Conf. Proceeding Series*, volume Part F1286, pages 167–174, 2017.
[2] Dzmitry Bahdanau et al. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
[3] Rakesh Bal et al. Analysing the extent of misinformation in cancer related tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 924–928, 2020.
[4] Colin R Blyth et al. Binomial confidence intervals. *Journal of the American Statistical Association*, 78(381):108–116, 1983.
[5] Célia Boyer and Ljiljana Dolamic. Automated detection of HONcode website conformity compared to manual detection: An evaluation. *J. of Medical Internet Research*, 17(6):e135, 2015.
[6] Lars Buitinck et al. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.
[7] Nitesh V Chawla et al. SMOTE: Synthetic Minority Over-ampling TEchnique. *J. of Artificial Intelligence Research*, 16:321–357, 2002.
[8] Wen-Ying Sylvia Chou et al. Addressing health-related misinformation on social media. *JAMA*, 320(23):2417–2418, 2018.
[9] Limeng Cui et al. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. *Proceedings of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 492–502, 2020.
[10] Arkin Dharawat et al. Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID-19 misinformation. *arXiv preprint arXiv:2010.08743*, 2020.
[11] Nicola Diviani et al. Exploring the role of health literacy in the evaluation of online health information: insights from a mixed-methods study. *Patient education and counseling*, 99(6):1017–1025, 2016.
[12] Gunther Eysenbach et al. From intermediation to disintermediation and apomediation: new models for consumers to access and assess the credibility of health information in the age of web 2.0. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics*. IOS Press, 2007.
[13] Yang Fan et al. Neural feedback text clustering with BiLSTM-CNN-Kmeans. *IEEE Access*, 6:57460–57469, 2018.
[14] Jian Feng et al. Web2Vec: Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning. *IEEE Access*, 8, 2020.
[15] Marcos Fernández-Pichel et al. Reliability prediction for health-related content: A replicability study. In *Proceedings of ECIR 2021, Lucca, Italy*, 2021.
[16] Sherry Girgis et al. Deep learning algorithms for detecting fake news in online text. In *Proceedings of ICCES 2018*, pages 93–97. IEEE, 2018.
[17] Anna Glazkova et al. g2tmn at Constraint@ AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection. *arXiv preprint arXiv:2012.11967*, 2020.
[18] Lorraine Goeuriot et al. Overview of the CLEF eHealth Evaluation Lab 2020. In *Int. Conf. of the Cross-Language Evaluation Forum for European Languages*, pages 255–271. Springer, 2020.
[19] Hema Karande et al. Stance detection with bert embeddings for credibility analysis of information on social media. *PeerJ Computer Science*, 7:e467, 2021.
[20] Yeolib Kim. Trust in health information websites: A systematic literature review on the antecedents of trust. *Health Informatics Journal*, 22(2):355–369, 2016.
[21] Laura Kinkead et al. Autodiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks. *BMC Medical Informatics and Decision Making*, 20(1):1–13, 2020.
[22] Jinhyuk Lee et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
[23] Pooja Malhotra et al. Breast cancer knowledge on line portal: an intelligent decision support system perspective. In *Australasian Conf. on Information Systems 2003*, pages 1–11. Edith Cowan University, 2003.
[24] Christine Marton. How women with mental health conditions evaluate the quality of information on mental health web sites: a qualitative approach. *J. of Hospital Librarianship*, 10(3):235–250, 2010.
[25] Corine S Meppelink et al. Reliable or not? an automated classification of webpages about early childhood vaccination using supervised machine learning. *Patient Education and Counseling*, 2020.
[26] Miriam J Metzger et al. Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Annals of the Int. Communication Association*, 27(1):293–335, 2003.
[27] Xing Pan et al. A review of cognitive models in human reliability analysis. *Quality and Reliability Engineering Int.*, 33(7):1299–1316, 2017.
[28] Nidhi A Patel and Rakesh Patel. A survey on fake review detection using machine learning techniques. In *4th Int. Conf. on Computing Communication and Automation (ICCCA)*, pages 1–6. IEEE, 2018.
[29] Fay Cobb Payton et al. Online HIV prevention information. *Internet Research*, 2014.
[30] Kashyap Popat et al. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM Int. on Conf. on Information and Knowledge Management*, pages 2173–2178, 2016.
[31] Hamman Samuel and Osmar Zaïane. Medfact: Towards improving veracity of medical information in social media using applied machine learning. In *Lecture Notes in Computer Science*, volume 10832, pages 108–120, 2018.
[32] Laura Sbaffi and Jennifer Rowley. Trust and credibility in web-based health information: a review and agenda for future research. *J. of Medical Internet Research*, 19(6):e218, 2017.
[33] Arabella Scantlebury et al. Experiences, practices and barriers to accessing health information: A qualitative study. *Int. J. of Medical Informatics*, 103:103–108, 2017.
[34] Julia Schwarz and Meredith Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 1245–1254, 2011.
[35] William M Silberg et al. Assessing, controlling, and assuring the quality of medical information on the internet: Caveant lector et viewor—let the reader and viewer beware. *JAMA*, 277(15):1244–1245, 1997.
[36] Parikshit Sondhi et al. Reliability prediction of webpages in the medical domain. In *Proceedings of ECIR 2012*, pages 219–231. Springer, 2012.
[37] Shijie others Song. The role of health literacy on credibility judgment of online health misinformation. In *2019 IEEE Int. Conf. on Healthcare Informatics (ICHI)*, pages 1–3. IEEE, 2019.
[38] Kristine Sørensen et al. Health literacy in europe: comparative results of the european health literacy survey (hls-eu). *European J. of Public Health*, 25(6):1053–1058, 2015.
[39] Hanna Suominen et al. Overview of the CLEF eHealth Evaluation Lab 2018. In *Int. Conf. of the Cross-Language Evaluation Forum for European Languages*, pages 286–301. Springer, 2018.
[40] Marco Viviani and Gabriella Pasi. Credibility in social media: opinions, news, and health information—a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5):e1209, 2017.
[41] Peter Williams et al. Health information on the internet: a qualitative study of nhs direct online users. In *Aslib proceedings*. MCB UP Ltd, 2003.
[42] Jue Xie. Sustaining quality assessment processes in user-centred health information portals. *AMCIS 2009 Proceedings*, page 189, 2009.
[43] Jue Xie and Frada Burstein. Using machine learning to support resource quality assessment: An adaptive attribute-based approach for health information portals. *Lecture Notes in Computer Science*, 6637:526–537, 2011.
[44] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40, 2020.