





Comparing Traditional and Neural Approaches for detecting Health-related Misinformation

Marcos Fernández-Pichel¹ , David E. Losada¹ , Juan C. Pichel¹ , and David Elswailer² 

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

Abstract. Detecting health-related misinformation is a research challenge that has recently received increasing attention. Helping people to find credible and accurate health information on the Web remains an open research issue as has been highlighted during the COVID-19 pandemic. However, in such scenarios, it is often critical to detect misinformation quickly [34], which implies working with little data, at least at the beginning of the spread of such information. In this work, we present a comparison between different automatic approaches of identifying misinformation, and we compare how they behave for different tasks and with limited training data. We experiment with traditional algorithms, such as SVMs or KNNs, as well as newer BERT-based models [5]. Our experiments utilise the CLEF 2018 Consumer Health Search task dataset [16] to perform experiments on detecting untrustworthy contents and information that is difficult to read. Our results suggest that traditional models are still a strong baseline for these challenging tasks. In the absence of substantive training data, classical approaches tend to outperform BERT-based models.

Keywords: Health-related content · Misinformation · Language · Neural approaches.

1 Introduction

The everyday use of the Web and social media has resulted in increased information accessibility [28]. The quality of information acquired via these channels is not assured, however, and infodemics with unreliable [1], inaccurate [6], or poor quality [29] information have become more common. Previous research has evidenced that providing poor quality search results in this context, leads people to make incorrect decisions [27]. People are influenced by search engine results and interacting with incorrect information results in poor choices being made.

Since search engines are widely used as a mean to find health advice online [8], misinformation provided via these services can be especially damaging, and there is a need to develop retrieval methods that can find trustworthy, and understandable search results. The quest for such high quality retrieval results was the primary goal of evaluation campaigns such as the CLEF Consumer Health Search task [16]. The urgent need for effective quality filtering devices has only been underlined during the 2020 pandemic, when large quantities of information about COVID-19 and its treatments was of questionable or poor quality [15,26]. Moreover, the early detection of health-related misinformation is critical to avoid potential personal injury [34]. This leads us to a scenario in which prediction must be based on low training data.

The evidence suggests that language is a good indicator to discern trustworthy from untrustworthy information [22]. Information of varying quality tends to differ in writing style and in the use of certain words [25]. For example, the use of technical terms or certain formalisms is associated with documents of higher quality and, in many cases, more trustful. Moreover, several machine learning technologies have been used to exploit linguistic properties of text [2,33].

In this work, we evaluate the performance of traditional classification approaches, such as SVMs or KNNs, and newer BERT-based models for detecting health-related misinformation. To that end, we employed the CLEF 2018 Consumer Health Search task dataset. This task focuses on providing high-quality health-related search results to non-expert users. Different experiments were performed using target variables such as trustworthiness, readability, and the combination of both. Following Hahnel et al. [12], we consider that for a document to be useful it should not only be trustful but also understandable by non-expert users.

The main objective of our research is to provide a thorough comparison between recent deep Natural Language Processing (NLP) models and traditional algorithms for the identification of poor quality online contents (untrustworthy and difficult to read web pages). We pay special attention to the behaviour of the models under realistic conditions (low training data). To that end, our study includes a report on the influence of the amount of training data in the effectiveness and the training time of the different models.

2 Related work

Several studies have analyzed how the credibility of online content is assessed [7,24,36]. Some interesting conclusions are that subjective ratings depend on the user's background, like years of education or reading skills [12]. Ginsca and colleagues [10] presented a thorough survey on existing credibility models from different information seeking perspectives.

Other researches focused on determining how the search engine result page (SERP) listings are used to determine credibility through user studies [18] or on the association between different features and reliability. For example, Grif-

fiths et al. [11] showed that algorithms like PageRank were unable to determine reliability on their own.

More specifically, some teams focused on assessing the credibility of health-related content on the web. For example, Matthews et al. [23] analysed a corpus about alternative cancer treatments and found that almost 90% contained false claims. Liao and Fu [19] studied the influence of age differences in credibility judgments and argued that older adults care less about the content of the site. Other teams focused on how to present medical information on a search engine result page to improve credibility judgments [31].

Sondhi and his colleagues presented an automatic approach, based on traditional learning algorithms, for medical reliability prediction at a document-level [33]. Other studies [37] proposed features, such as those based on sentiment or polarity signals, to better detect misinformation.

Recent advances have shown that new neural approaches can be effective tools for detecting health-related misinformation [4,9,14,32]. Most of these methods employ not only content-based features but other signals (e.g. network-based features).

In this work, we present an innovative comparison between traditional learning methods, such as SVMs or KNNs, and neural approaches for identifying health misinformation. We also test how the models behave with low training data, and our study is constrained to work with models that are fed with content-based features.

3 Dataset

To perform this comparison, we selected the CLEF 2018 Consumer Health Search task dataset [16], which focuses on the effectiveness of health-related information provided by search engines. The search task aims at helping non-expert users who are looking for health-advice. The dataset contains webpages obtained from CommonCrawl³. The creators of the dataset 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 manually extended by adding sites known to be either trustful or untrustful.

The assessments were provided by human assessors from Amazon Mechanical Turk. The turkers labelled the documents with respect to three different query-dependent dimensions: relevance, trustworthiness, and readability. In our experiments we consider only the latter two.

Both dimensions of interest were judged on an eleven point scale, from 0 to 10. In our case, we wanted to approach the problem as a two-class classification challenge and, thus, we converted the original scores into binary variables. To that end, we removed the middle values (from 4 to 6) and mapped the extreme values to trustful/untrustful and readable/non-readable respectively. Table 1 reports the main statistics of the resulting datasets. We also tested classifiers for

³ <http://commoncrawl.org/>

	Trustworthiness	Readability	Useful (T&R)
# Positive	10,405	3,102	1,567
% Positive	73%	20%	12%
# Negative	3,820	12,455	11,488
% Negative	27%	80%	88%

Table 1. Label distribution in the CLEF eHealth dataset.

the task of distinguishing between *useful* documents for non-expert end users (i.e., trustworthy and readable) and *non-useful* documents (the remaining documents). With this goal in mind, we labelled useful documents as those that are both trustworthy and readable (third column in the table).

4 Experimental Design

The experiments were conceived such that the aim was to uncover misinformative documents, as measured by the dimensions considered: trustworthiness, readability, and the combination of both. To that end, we compared the performance of traditional models against BERT models.

We employed a 5-fold stratified cross-validation strategy in all the experiments. To address the imbalance in data labels, we also applied a cost-factor strategy [13,21] in those learning methods whose implementation supports it⁴. We decided to set this cost-factor to the proportion between the classes for each experiment.

All experiments were conducted using the same docker container environment, an image with Ubuntu 18.04 and Python 3.7.3 version. The host machine also had 32GB of RAM, 240GB of storage, an Intel(R) Core(TM) i7-9750H CPU @ 1.60GHz, and a Nvidia Tesla V100S 32GB GPU, which was beneficial for the BERT experiments.

4.1 Traditional models

We employed two variants for these experiments. The first consisted of a model where each word in a document was considered as a different feature, weighted by its normalized frequency. The second was equivalent, but stopwords were removed. The vocabulary was pruned to only consider terms present in at least 10% of the training corpus in both variants. We also applied a standardisation of the features (to get 0 mean and 1 standard deviation).

- **SVM.** Following [33], a classic reference for health information reliability detection, we used a support vector machine implemented as part of the SVMlight toolkit [17]⁵.

⁴ We employed <https://scikit-learn.org/stable/> (version 0.24.1)

⁵ Using default parameter setting (kernel linear and $C = [avg. x * x]^{-1}$). We employed the SVMlight Python wrapper with this configuration.

- **Random Forest (RF)**. We used Random Forest scikit-learn default implementation (100 trees were used and the Gini index was the criterion to measure the quality of a split).
- **Naive Bayes (NB)**. We used Naive Bayes scikit-learn default implementation, utilising the Multinomial Bayes variant, which is particularly recommended for imbalanced data problems.
- **KNN**. We used scikit-learn default implementation of the KNN classifier ($k = 5$ neighbours).

For the models whose implementation supports cost weighting (SVM and RF) we also ran experiments with cost-weighting variants⁶.

4.2 BERT-based models

For neural approaches, we considered BERT-based models [5]. These are pre-trained neural networks based on transformers architecture, and lead to state-of-the-art solutions for many NLP tasks.

More specifically, we used **DistilBERT base** model (uncased version) [30] and **DistilRoBERTa base** model from HuggingFace Transformers library [35]. The first has 6 layers, 768 hidden, 12 heads, and 66M parameters, while the second has the same number of layers, hidden and heads, but 82M parameters. These are light models obtained from larger ones, such as BERT base [5] or RoBERTa base [20]. The distilled models reduce the number of layers by a factor of 2, and the number of parameters by 40% while retaining 97% of the original performance [30].

These models were fine-tuned for our task in each fold. For the training process, 4 epochs and a 10% validation split were used, with a learning rate of 2^{-5} , a training batch size of 32, and a validation batch size of 64 instances.

We note that BERT models have an input limit of 512 tokens. This was a challenge since the majority of the documents were larger. We trained the models with the first 512 tokens of each training document. At testing time, two different approaches were evaluated: i) making the prediction using only the first 512 tokens of the test document, or ii) segmenting each test document into 512-token chunks, passing the classifier on each chunk, and returning a final score that is the prediction score averaged over all chunks (aggregation strategy). Both strategies are reported and compared in Section 5.

5 Experimental Results

A set of experiments was performed for each target classification problem. We report the results for each of the different dimensions and models, providing the F1-score (harmonic mean between precision and recall) for each class and the macro average F1 (unweighted mean of F1-score per class).

	Cost factor	F1 macro	F1 trustful	F1 untrustful
SVM (stopword removal)	1	0.57	0.84	0.3
SVM	1	0.57	0.83	0.31
SVM n-grams (stopword removal)	1	0.57	0.84	0.29
SVM n-grams	1	0.57	0.84	0.3
RF (stopword removal)	1	0.57	0.84	0.3
RF	1	0.57	0.84	0.29
Naive Bayes (stopword removal)	1	0.59	0.76	0.41
Naive Bayes	1	0.59	0.78	0.39
KNN (stopword removal)	1	0.6	0.8	0.39
KNN	1	0.59	0.82	0.36
DistilBERT	1	0.61	0.82	0.39
DistilRoBERTa	1	0.59	0.82	0.36
DistilBERT (aggregation)	1	0.58	0.83	0.33
DistilRoBERTa (aggregation)	1	0.61	0.84	0.38
<hr/>				
SVM (stopword removal)	2.72	0.56	0.7	0.42
SVM	2.72	0.57	0.71	0.42
SVM n-grams (stopword removal)	2.72	0.57	0.71	0.43
SVM n-grams	2.72	0.57	0.71	0.43
RF (stopword removal)	2.72	0.57	0.84	0.29
RF	2.72	0.56	0.84	0.27
DistilBERT	2.72	0.6	0.74	0.45
DistilRoBERTa	2.72	0.59	0.72	0.46
DistilBERT (aggregation)	2.72	0.57	0.69	0.45
DistilRoBERTa (aggregation)	2.72	0.58	0.7	0.46

Table 2. Trustworthiness results obtained when setting or not the cost-factor to the proportion between classes.

5.1 Trustworthiness

The first dimension considered was trustworthiness. For this task, there is no substantial difference between the models (see Table 2). KNN and NB seem to be slightly superior to the other classic models and comparable to the best BERT-based variants.

With cost-weighting settings, the models tend to improve the detection of the minority class (untrustful), but the relative merits of the models remain essentially the same. Only RF shows here a distinctive behaviour, as its cost-weight variant decreases performance in terms of F1 untrustful.

Stopword removal had no substantial effect and the use of n-grams (bigrams and trigrams) did not bring any improvement (that is why it is only reported for SVMs). On the other hand, the aggregation strategy for BERT models did not yield any substantial advantage over a prediction that is solely based on the leading chunk. Making predictions with a single chunk of the test document is computationally convenient, and our experiments suggest that this approach is comparable to a more thorough prediction based on the entire test document.

Overall, these results suggest that BERT models are unable to improve over simpler (and computationally less expensive) approaches. This could be related

⁶ Scikit-learn does not support cost-weighting for NB and KNN.

	Cost factor	F1 macro	F1 readable	F1 non-readable
SVM (stopword removal)	1	0.5	0.13	0.86
SVM	1	0.49	0.12	0.86
SVM n-grams (stopword removal)	1	0.49	0.11	0.86
SVM n-grams	1	0.49	0.12	0.86
RF (stopword removal)	1	0.51	0.16	0.86
RF	1	0.51	0.16	0.86
Naive Bayes (stopword removal)	1	0.59	0.33	0.84
Naive Bayes	1	0.59	0.33	0.84
KNN (stopword removal)	1	0.52	0.21	0.82
KNN	1	0.52	0.2	0.83
DistilBERT	1	0.5	0.19	0.81
DistilRoBERTa	1	0.49	0.16	0.81
DistilBERT (aggregation)	1	0.51	0.2	0.82
DistilRoBERTa (aggregation)	1	0.49	0.15	0.82
SVM (stopword removal)	4.02	0.51	0.3	0.72
SVM	4.02	0.5	0.31	0.68
SVM n-grams (stopword removal)	4.02	0.52	0.32	0.72
SVM n-grams	4.02	0.52	0.33	0.71
RF (stopword removal)	4.02	0.52	0.17	0.86
RF	4.02	0.53	0.18	0.87
DistilBERT	4.02	0.47	0.27	0.67
DistilRoBERTa	4.02	0.5	0.3	0.69
DistilBERT (aggregation)	4.02	0.49	0.28	0.7
DistilRoBERTa (aggregation)	4.02	0.48	0.27	0.69

Table 3. Readability results obtained when setting or not the cost-factor to the proportion between classes.

to the lack of large amounts of training data. In Section 5.4, we further analyze the models under varying training sizes.

5.2 Readability

In the readability experiments the objective was to detect the documents labelled as non-readable from the collection. The results in the readability experiments (see Table 3) show that the traditional algorithms perform better than BERT models. In particular, Naive Bayes achieves the best performance overall. When we set the *cost-factor* = 4.02 (notice that in this case the majority class was the non-readable), conclusions remain the same. Again, removing stopwords had no substantial effect on performance and the BERT-based models do not benefit from the aggregation approach.

These results suggest that determining readability can be effectively addressed with standard word-based technology. Even a simple bag-of-words model using a traditional learning method (like Naive Bayes or KNN) forms a solid classifier, comparable to the best neural models. One could argue that readability classification is essentially about distinguishing between the usage of simpler vs complex language. Our experiments show that such a goal can be competently tackled by classic NB technology.

	Cost factor	F1 macro	F1 useful docs	F1 non-useful docs
SVM (stopword removal)	1	0.51	0.1	0.92
SVM	1	0.5	0.07	0.93
SVM n-grams (stopword removal)	1	0.51	0.09	0.93
SVM n-grams	1	0.5	0.07	0.93
RF (stopword removal)	1	0.5	0.07	0.92
RF	1	0.5	0.06	0.93
Naive Bayes (stopword removal)	1	0.59	0.3	0.88
Naive Bayes	1	0.6	0.32	0.88
KNN (stopword removal)	1	0.54	0.16	0.92
KNN	1	0.53	0.15	0.91
DistilBERT	1	0.56	0.2	0.91
DistilRoBERTa	1	0.53	0.12	0.93
DistilBERT (aggregation)	1	0.56	0.2	0.91
DistilRoBERTa (aggregation)	1	0.54	0.16	0.91
<hr/>				
SVM (stopword removal)	7.33	0.57	0.3	0.84
SVM	7.33	0.54	0.29	0.79
SVM n-grams (stopword removal)	7.33	0.58	0.31	0.84
SVM n-grams	7.33	0.55	0.3	0.8
RF (stopword removal)	7.33	0.51	0.1	0.92
RF	7.33	0.51	0.09	0.92
DistilBERT	7.33	0.57	0.29	0.84
DistilRoBERTa	7.33	0.5	0.27	0.73
DistilBERT (aggregation)	7.33	0.55	0.29	0.81
DistilRoBERTa (aggregation)	7.33	0.49	0.26	0.72

Table 4. Usefulness results obtained when setting or not the cost-factor to the proportion between classes.

5.3 Usefulness (trustworthiness & readability)

We also performed experiments combining readability and trustworthiness. To that end, we considered as *useful* documents the ones labelled as both trustful and readable. This seems reasonable since non-expert users look for trustworthy and understandable health-advice on the Web [12]. The remaining documents are regarded as *non-useful* documents (highly technical or untrustful).

The results (see Table 4) suggest that, as was the case in the trustworthiness experiments, there is no substantial difference between traditional and BERT models. Only a slight improvement of Naive Bayes over the rest was found. Again, applying a cost-sensitive learning strategy, improves the minority class detection, but RF does not benefit from this technique.

5.4 Influence of the training set size

In order to evaluate the influence of the training set size on effectiveness and efficiency, we report here two experiments: one for trustworthiness and another one for readability.

We selected **Naive Bayes**, **KNN**, and **DistilBERT base** (keeping stopwords and without any cost-factor), which were the best performing models in the experiments reported above. A 5-fold cross-validation strategy was applied

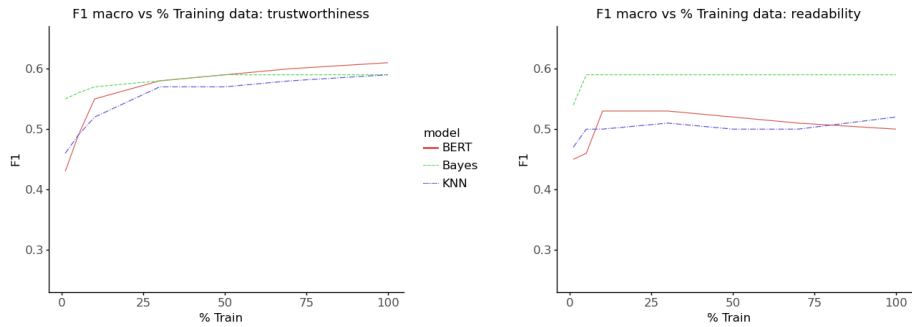


Fig. 1. Variation of the F1 macro precision with percent training data used in trustworthiness and readability tasks.

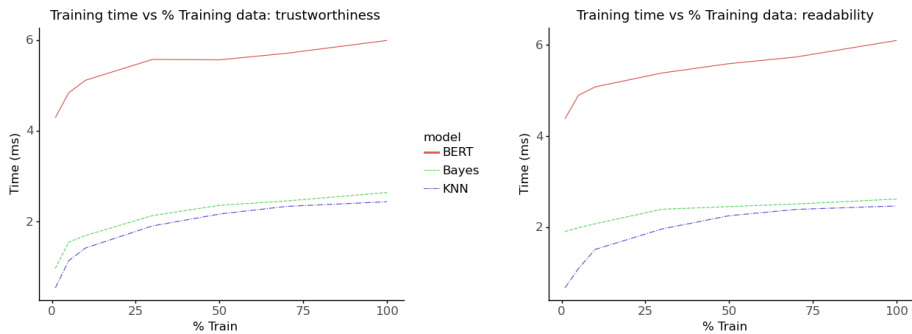


Fig. 2. Variation of the training time (ms) with percent training data used in trustworthiness and readability tasks. *Y* axis in log scale.

again, but in this case models were only trained using a percentage of the training fold (always ensuring a stratified sample). We considered 1%, 5%, 10%, 30%, 50%, 70%, and 100% of the available data.

In Figure 1, we depict how the F1 macro-precision of each model evolves with varying training data sizes. For trustworthiness (graph on the left), Naive Bayes clearly outperforms DistilBERT and KNN when training data is scarce. However, as we inject more training data, the performance of NB flattens, while the other models tend to benefit from the availability of more training examples. With the full training set, the three models perform roughly the same but the graph suggests that KNN and DistilBERT would keep improving and eventually beat the NB classifier.

For readability (graph on the right), Naive Bayes is the best performer over all training sizes. However, all models perform well even with few training examples. This supports the claim that few examples suffice to build a readability classifier. Observe that the performance of the three models tends to flatten (or even gets worse) with more than 20% of the training examples.

In Figure 2, we report the training times required by each model against the percentage of the training data. In both tasks, the training time taken by DistilBERT is much longer than that taken by the other models (we had to use a logarithmic scale for the representation). KNN is faster than Naive Bayes since it is a *lazy* approach (in training time it only stores the examples and learns no model).

Finally, we also computed the prediction time (time needed to classify a test instance). On average, Naive Bayes took 4.9 μ s to predict, KNN 300 μ s, and DistilBERT 0.002 μ s. These results make sense since KNN has higher computational load in prediction time (needs to search for the neighbours). The DistilBERT model shows a surprisingly low average time, which could be due to the fact that the underlying library is very optimized and takes advantage of the host GPU, while traditional models are only set to be executed in CPU.

6 Conclusions and Future Work

In this work, we presented a comparison between traditional learning methods, such as SVMs or KNNs, and neural approaches such as BERT models, for automatically identifying health-related misinformation online. We also tested how they behave with varying sizes of training data. The main lesson extracted from the study is that, for these tasks and dataset, the added complexity of a neural model does not seem to be worthwhile. Sophisticated neural models were outperformed here by traditional models and the advantage of these classic methods is even more apparent with small training sets.

The results are modest overall and there is still room for improvement, as the tasks are difficult and more research effort is required. The main conclusion is that a traditional model such as NB is consistent (with very different sizes of training data), is computationally efficient and should not be discarded considering that in many environments we have little training data.

This study opens up new lines of research related to how to detect health-related misinformation on the Web. A natural next step could be testing other strategies to deal with BERT input limit, such as generating summaries of the test documents and, subsequently predicting based on the summaries or, alternatively, using neural models that have no input limit, such as LongFormer [3].

Finally, we could also consider extending these experiments with BERT models already fine tuned for a document classification task.

References

1. 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)
2. Adhikari, A., Ram, A., Tang, R., Lin, J.: Docbert: Bert for document classification. arXiv preprint arXiv:1904.08398 (2019)

3. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150 (2020)
4. Cui, L., Lee, D.: Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885 (2020)
5. 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)
6. Eysenbach, G.: Infodemiology: The epidemiology of (mis) information. *The American journal of medicine* **113**(9), 763–765 (2002)
7. Fogg, B.J.: Prominence-interpretation theory: Explaining how people assess credibility online. In: CHI'03 extended abstracts on human factors in computing systems. pp. 722–723 (2003)
8. Fox, S.: Health topics: 80% of internet users look for health information online. Pew Internet & American Life Project (2011)
9. Giachanou, A., Rosso, P., Crestani, F.: Leveraging emotional signals for credibility detection. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 877–880 (2019)
10. Ginsca, A.L., Popescu, A., Lupu, M.: Credibility in information retrieval. *Found. Trends Inf. Retr.* **9**(5), 355–475 (Dec 2015). <https://doi.org/10.1561/1500000046>, <https://doi.org/10.1561/1500000046>
11. Griffiths, K.M., Tang, T.T., Hawking, D., Christensen, H.: Automated assessment of the quality of depression websites. *Journal of Medical Internet Research* **7**(5), e59 (2005)
12. Hahnel, C., Goldhammer, F., Kröhne, U., Naumann, J.: The role of reading skills in the evaluation of online information gathered from search engine environments. *Computers in Human Behavior* **78**, 223–234 (2018)
13. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications* **73**, 220–239 (2017)
14. Hossain, T., Logan IV, R.L., Ugarte, A., Matsubara, Y., Singh, S., Young, S.: Detecting covid-19 misinformation on social media. In: Workshop on natural language processing for COVID-19 (NLP-COVID) (2020)
15. Islam, M.S., Sarkar, T., et al.: Covid-19-related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene* **103**(4), 1621–1629 (2020)
16. Jimmy, J., Zuccon, G., Palotti, J., Goeriot, 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)
17. Joachims, T.: Making large-scale support vector machine learning practical. In: *Advances in kernel methods: support vector learning*, pp. 169–184 (1999)
18. Kattenbeck, M., Elsweiler, D.: Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management* (2019)
19. Liao, Q.V., Fu, W.T.: Age differences in credibility judgments of online health information. *ACM Transactions on Computer-Human Interaction (TOCHI)* **21**(1), 1–23 (2014)
20. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. *CoRR abs/1907.11692* (2019), <http://arxiv.org/abs/1907.11692>
21. Madabushi, H.T., Kochkina, E., Castelle, M.: Cost-sensitive bert for generalisable sentence classification with imbalanced data. arXiv preprint arXiv:2003.11563 (2020)

22. Matsumoto, D., Hwang, H.C., Sandoval, V.A.: Cross-language applicability of linguistic features associated with veracity and deception. *Journal of Police and Criminal Psychology* **30**(4), 229–241 (2015)
23. 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)
24. 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)
25. 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)
26. 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. *Psychological science* **31**(7), 770–780 (2020)
27. 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)
28. Reuters Insitute, University of Oxford: *Reuters Digital News Report 2020* (2020 (accessed November 16, 2020)), <https://www.digitalnewsreport.org/survey/2020>
29. Rieh, S.Y.: Judgment of information quality and cognitive authority in the web. *Journal of the American society for information science and technology* **53**(2), 145–161 (2002)
30. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv abs/1910.01108* (2019)
31. 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)
32. Sicilia, R., Giudice, S.L., Pei, Y., Pechenizkiy, M., Soda, P.: Twitter rumour detection in the health domain. *Expert Systems with Applications* **110**, 33–40 (2018)
33. Sondhi, P., Vydiswaran, V.V., Zhai, C.: Reliability prediction of webpages in the medical domain. In: *European Conference on Information Retrieval*. pp. 219–231. Springer (2012)
34. Vigdor, N.: Man fatally poisons himself while self-medicating for coronavirus, doctor says (March 2020), <https://www.nytimes.com/2020/03/24/us/chloroquine-poisoning-coronavirus.html>, [Online; posted 24-March-2020]
35. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
36. 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)
37. Zhao, Y., Da, J., Yan, J.: Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. *Information Processing & Management* **58**(1), 102390 (2021)