



Rzeczpospolita
Polska

Unia Europejska
Europejski Fundusz Społeczny



POLISH-JAPANESE ACADEMY
OF INFORMATION TECHNOLOGY

Informatics

mgr inż. Aleksandra Nabożny

**Credibility Evaluation
of Online Health Information
using Human in the Loop
Machine Learning**

Ph.D. Thesis

Supervisor:

prof. dr hab. Adam Wierzbicki

Auxiliary supervisor:

dr Konrad Maj

March 2023

Acknowledgments

*Wszystkim współautorom oraz
mojej wspaniałej rodzinie.*

Abstract

Evaluating the credibility of medical content on the Internet is becoming increasingly urgent in the 21st century. However, countless amounts of data published daily online do not allow for manual evaluation of their content by domain experts. On the other hand, decisions based on false medical recommendations can be so severe that the final classification of credibility ought to be made by a human.

This doctoral dissertation describes work to improve the process of evaluating Online Health Information by experts. This work takes the essential steps toward creating an expert-supported, semi-automated system for capturing and tagging unreliable medical texts appearing on the Web.

Three experiments were carried out to collect the necessary data (medical content with expert assessments). They were followed by four analyses, each described in a separate article, all part of this dissertation. The first experiment evaluated single sentences in two modes - with and without knowing the context of the entire article. The results of this analysis indicated the great difficulty that experts experienced assessing sentences without context, and therefore a second experiment was carried out. It tested four different methods for enriching the context of a single sentence. As a result, an efficient unit of text was defined for the evaluated content. It consists of three consecutive sentences with keywords.

The first article describes the two experiments and data analysis mentioned above. The second article describes the third experiment, which aimed to create a dataset in which selected text units extracted from online medical articles were evaluated by domain experts (psychiatry, gynecology, cardiology, and pediatrics). The obtained data was open-sourced. The second article also describes an analysis that detects rhetorical patterns that mislead experts, distorting their credibility as-

assessment. The third article presents filtering classifiers created to maximize the efficiency of an expert working on annotation. The fourth article concerns the study of the explanatory capabilities of the results returned by filtering classifiers.

The results of the qualitative analysis indicate the existence of repetitive rhetorical patterns that appear in non-credible medical content. Schemes similar to those recognized in the general domain of disinformation and specific to popular science medical content can be identified. Classifiers allow for pre-filtration, which accelerates twice the detection of unreliable content by the annotating expert. The explanatory capabilities of classifiers depend on the degree of compression of the input data. Better generalization of results (applying the same classifiers to broader topics) prevents insight into decisions related to semantic attributes. In comparison, minor generalization allows for it but requires constructing separate classifiers for thematically narrow domains.

A theoretical system in which a sufficiently large group of experts would evaluate all data published on the Internet in real-time is impossible to implement. Therefore, the efforts focused on maximizing the throughput of the expert-supported assessment system. The throughput was improved two-fold. Firstly, the results of the experiments allowed for the isolation of fragments of medical texts - three sentences. They are small enough to be a meaningful unit for crowd-sourced data collection. At the same time, they are complex enough so that the expert evaluator retains the context needed for the assessment. Secondly, an expert can catch twice as many unreliable examples using the created filtering classifiers.

In addition, analyzing the filtering algorithms allows for selecting such parameters to obtain the desired feedback for the end user.

The qualitative analysis of the obtained credibility labels indicates that cognitive biases, to some extent, distort the

medical expert assessment. These conclusions define new research directions in the psychology of disinformation required to create the system mentioned above.

Streszczenie

Ocena wiarygodności treści medycznych pojawiających się w Internecie staje się w XXI wieku coraz bardziej palącą potrzebą. Z jednej strony, niezliczone ilości danych publikowanych codziennie online nie pozwalają na ręczną ocenę treści przez ekspertów dziedzinowych. Z drugiej strony, decyzje podejmowane w oparciu o fałszywe zalecenia medyczne mogą być na tyle poważne w skutkach, że ostateczną klasyfikację wiarygodności powinien podejmować człowiek.

Niniejsza rozprawa doktorska zawiera opis prac mających na celu usprawnienie procesu oceny medycznych treści internetowych przez ekspertów. Prace te stanowią przygotowanie do stworzenia wspieranego przez ekspertów, półautomatycznego systemu wychwytywania i tagowania niewiarygodnych stwierdzeń medycznych pojawiających się w Sieci. Przeprowadzono trzy eksperymenty, których celem było zebranie danych (treści medyczne wraz z ocenami eksperckimi). Następnie przeprowadzono cztery analizy, każda opisana w oddzielnym artykule. Artykuły stanowią załączniki do niniejszej rozprawy. Pierwszy eksperyment polegał na ocenie pojedynczych zdań w dwóch trybach - z i bez znajomości kontekstu całego artykułu. Wyniki analizy wykazały dużą trudność ekspertów w ocenie zdań bez kontekstu, w związku z czym przeprowadzono drugi eksperyment. W drugim eksperymencie przetestowano cztery różne metody wzbogacania kontekstu pojedynczego zdania. W jego wyniku zdefiniowana została jednostka tekstowa dla ocenianych treści - trójka zdań wraz ze słowami kluczowymi. Pierwszy artykuł stanowi opis dwóch wyżej wymienionych eksperymentów oraz analizy danych. W drugim artykule opisano trzeci eksperyment - stworzenie zbioru danych, w którym wybrane jednostki tekstowe, wyodrębnione z popularnonaukowych artykułów medycznych, ocenione zostały przez ekspertów dziedzinowych (psychiatria, ginekologia, kar-

diologia i pediatria). Dane udostępniono. Drugi artykuł opisuje również analizę polegającą na wykryciu schematów retorycznych, które wprowadzają ekspertów w błąd, wypaczając ich ocenę wiarygodności. Trzeci artykuł opisuje stworzone klasyfikatory filtrujące, mające na celu maksymalizację efektywności eksperta pracującego przy anotacji. Czwarty artykuł opisuje badanie możliwości interpretacyjnych wyników zwracanych przez algorytmy filtrujące.

Wyniki analizy jakościowej wykazują istnienie powtarzalnych schematów retorycznych, które pojawiają się w niewiarygodnych treściach medycznych. Istnieją schematy zarówno podobne do tych rozpoznawanych w dezinformacji pozadziennowej, jak i specyficzne dla popularnonaukowych treści medycznych. Wykorzystanie klasyfikatorów pozwala na wstępną filtrację, która przyspiesza wykrywanie treści niewiarygodnych przez anotującego eksperta średnio dwukrotnie, w zależności od dziedziny. Możliwości objaśniające klasyfikatorów zależą od stopnia kompresji danych wejściowych. Lepsza generalizacja wyników uniemożliwia wgląd w decyzje związane z atrybutami semantycznymi, podczas gdy mniejsza generalizacja to umożliwia, ale wymaga budowy oddzielnych klasyfikatorów dla wąskich tematycznie dziedzin.

Teoretyczny system, w którym odpowiednio szerokie grono ekspertów oceniałoby wszystkie dane publikowane w Internecie w czasie rzeczywistym w praktyce nie jest możliwy do wykonania. Dlatego moje wysiłki skupione były wokół maksymalizacji przepustowości wspieranego przez eksperta systemu ocen. Udało się uzyskać polepszenie przepustowości dwutorowo. Po pierwsze, wyniki eksperymentów pozwoliły wyodrębnić fragmenty tekstów medycznych - trójki zdań - które są na tyle małe, aby stanowiły sensowną jednostkę do zbierania danych za pomocą crowd-sourcingu, jednocześnie będąc na tyle złożone, aby ekspert oceniający nie tracił kontekstu potrzebnego do oceny. Po drugie, stworzone klasyfikatory fil-

trujące sprawiają, iż ekspert etykietujący tę samą liczbę treści co bez korzystania z filtracji, jest w stanie wyłapać średnio dwukrotnie więcej niewiarygodnych przykładów.

Ponadto, analiza algorytmów filtrujących pozwala na dobranie takich parametrów, aby uzyskać pożądaną informację zwrotną dla użytkownika końcowego.

Analiza jakościowa uzyskanych etykiet wiarygodności wskazuje na to, iż ekspercka ocena jest w pewnym stopniu wypaczana poprzez skrzywienia poznawcze. Wnioski te definiują nowe kierunki badań z zakresu psychologii dezinformacji, wymagane do stworzenia rzeczzonego systemu.

Contents

1 Introduction	1
1.1 Aim and scope	1
1.2 Research objectives	3
1.2.1 Defining an optimal unit for labeling online health information in terms of credibility.	5
1.2.2 Creating a dataset of selected medical tex- tual units labeled in terms of their credibility	5
1.2.3 Understanding the syntactic, semantic, and rhetorical structure of the new units of medical disinformation.	6
1.2.4 Investigating the relationship between ma- nipulation types in a medical domain and cognitive biases	6
1.2.5 Designing an annotation protocol for la- beling medical content for a given textual unit	6
1.2.6 Designing filtering methods for selected textual units	7
1.2.7 Interpreting and explaining decisions made by classification models	7
2 Literature Review	9

Contents

2.1	Source credibility	10
2.1.1	Source credibility in the general domain	10
2.1.2	Source credibility in the medical domain	12
2.2	Message credibility	13
2.2.1	Message credibility assessment support tools	14
2.2.2	Message credibility assessment in the medical domain	18
2.2.3	Language modeling as a support tool for message credibility assessment	24
2.2.4	Context enrichment for assessing short message credibility	26
2.3	Implications	27
3	Contributions	30
3.1	Contribution 1. Proposing sentence triplets as a unit of medical credibility evaluation.	30
3.2	Contribution 2. Creation of a dataset of 10,000 sentences from online medical articles with credibility evaluations made by medical experts using a new annotation protocol	32
3.3	Contribution 3. Identification and description of manipulation techniques prevalent in online health information content	34
3.4	Contribution 4. Topical classifiers of credibility of medical sentence triplets with 90% precision in credible class	38

Contents

3.5 Contribution 5. Comparison of two classification methods in terms of their generalization and explanatory power	39
4 Discussion	41
4.1 Concluding remarks	41
4.2 Limitations	42
4.3 Future work	43
References	48
5 Articles comprising the thesis	63
5.1 Article 1 "Enriching the Context: Methods of Improving the Non-contextual Assessment of Sentence Credibility" (WISE 2019, 140 pts)	63
5.2 Article 2 "Active Annotation in Evaluating the Credibility of Web-Based Medical Information: Guidelines for Creating Training Data Sets for Machine Learning" (Journal of Medical Internet Research, Medical Informatics, 70 pts.)	80
5.3 Article 3 "Focus On Misinformation: Improving Medical Experts' Efficiency Of Misinformation Detection" (WISE 2021, 140 pts., Awarded as 'Best Student Paper runner-up')	100

5.4 Article 4 **"Improving medical experts' efficiency of misinformation detection: an exploratory study"** (World Wide Web, 100 pts.) 116

CHAPTER 1

Introduction

1.1 Aim and scope

Increasing access to and use of the World Wide Web makes it a primary source of information for an ever-growing number of people worldwide¹. This phenomenon also includes learning about health, making people more likely to browse the Web seeking therapies for various medical conditions rather than consulting their case with a medical doctor². On the other hand, this means that the scale of the problems related to the spread of fake or out-of-date medical articles on the Web is increasing.

Thus, mechanisms enabling the detection and verification of online content likely to mislead potential readers are indispensable and can have a major and positive social impact. This doctoral dissertation aims to partially automate the evaluation of the credibility of the medical content published on the World Wide Web aimed at the general public, excluding expert materials for professionals.

Rapidly changing medical recommendations and discoveries make the task non-trivial. Something reliable published yesterday is not necessarily trustworthy today. The quality of the content degrades over time unless it is updated. Moreover, the evaluation itself may be complex. For example, the abovementioned content often

¹<https://www.zippia.com/advice/how-many-people-use-the-internet/>

²<https://www.wect.com/2019/06/24/study-finds-us-citizens-turn-google-before-their-doctor/>

Chapter 1. Introduction

consists of sentences that express factual statements but miss the critical context, draw incorrect conclusions from valid and adequately cited research, or exaggerate the real side effects of specific therapies. It is difficult to distinguish between the truthfulness and falsehood of a given claim without expert knowledge and experience. It should also be noted that experts and non-specialists are burdened with cognitive biases that either increase their susceptibility to misinformation or skew their credibility assessments. It should be noted that the term "misinformation" that is used throughout the dissertation differs from the more prevalent term "disinformation". Medical disinformation is information that is deliberately created to mislead and harm a patient, while medical misinformation is information that is misleading but not created with the intention of causing harm. Unlike in the general news domain, in the medical domain, false information is often disseminated without malicious intentions.

I assumed that the final decision regarding the credibility of the medical content should be made by a human - an expert practitioner of medicine. All my efforts were therefore directed toward creating a set of methods to optimize the experts' work in a potential crowd-sourced expert-in-the-loop credibility classification system. The existing research uses various approaches to entirely or partially automate medical credibility evaluations. However, none of these works have considered the question - what is an optimal unit of information for evaluating the credibility of information by medical experts? I started this work by considering the above question. I have proposed a new unit of information and designed methods for supporting credibility evaluations that are best suited for this unit and do not require additional information. The resulting methods can effectively support medical experts who can focus on evaluating the credibility of medical content using evidence-based medicine. My work fits into the broader context of the human-in-the-loop machine learning (HITL-ML) trend, as summarized and defined in [1]. The assumptions inherent in the functioning of my system can be defined as Interactive Machine Learning (IML) for the following reasons:

1. IML assumes that the human expert selects samples for learning the set; In my project, medical experts defined narrow thematic fields based on their own experience of medical practice.

2. Firebrink et al. [2] stated that in an IML system, the evaluation of the models should go beyond their accuracy and include subjective judgments. In this project, such assessments include the optimal use of expert time and the throughput of a continuous system.

I conducted several empirical studies with the participation of medical experts. In the first experiment, I collected preliminary credibility assessments of individual sentences extracted from articles following two scenarios. In the first scenario, experts had access to the context of the whole article, while in the second scenario, the sentences were assessed context-free. In the second experiment, I tested several methods of context enrichment for the individual sentences to decide on an optimal unit of information to perform a credibility assessment. I was looking for a unit that is as short as possible but contains enough context to be apt for assessment. For the second experiment, I used data from the first experiment. In the third experiment, I collected a dataset of over 10 000 sentences that were assessed with compressed context, which was obtained using the methods tested in the second experiment. The qualitative analysis of the dataset allowed for the identification of manipulation techniques and rhetorical structures prevalent in medical misinformation. In the fourth experiment, I created credibility classifiers that may serve as filtering methods for assessing the textual units. The filtering methods aimed to optimize an expert's time in the expert-in-the-loop credibility assessment system so that the more non-credible messages they got, the more efficient the system would be. The last experiment was performed to study the generalization ability and explainability of the designed classification models.

1.2 Research objectives

I have used seven research objectives to structure my research. Figure 1.1 shows the objectives, the experiments, and the resulting deliverables on the timeline, as well as the assignment of these works to the four research papers comprising this Thesis.

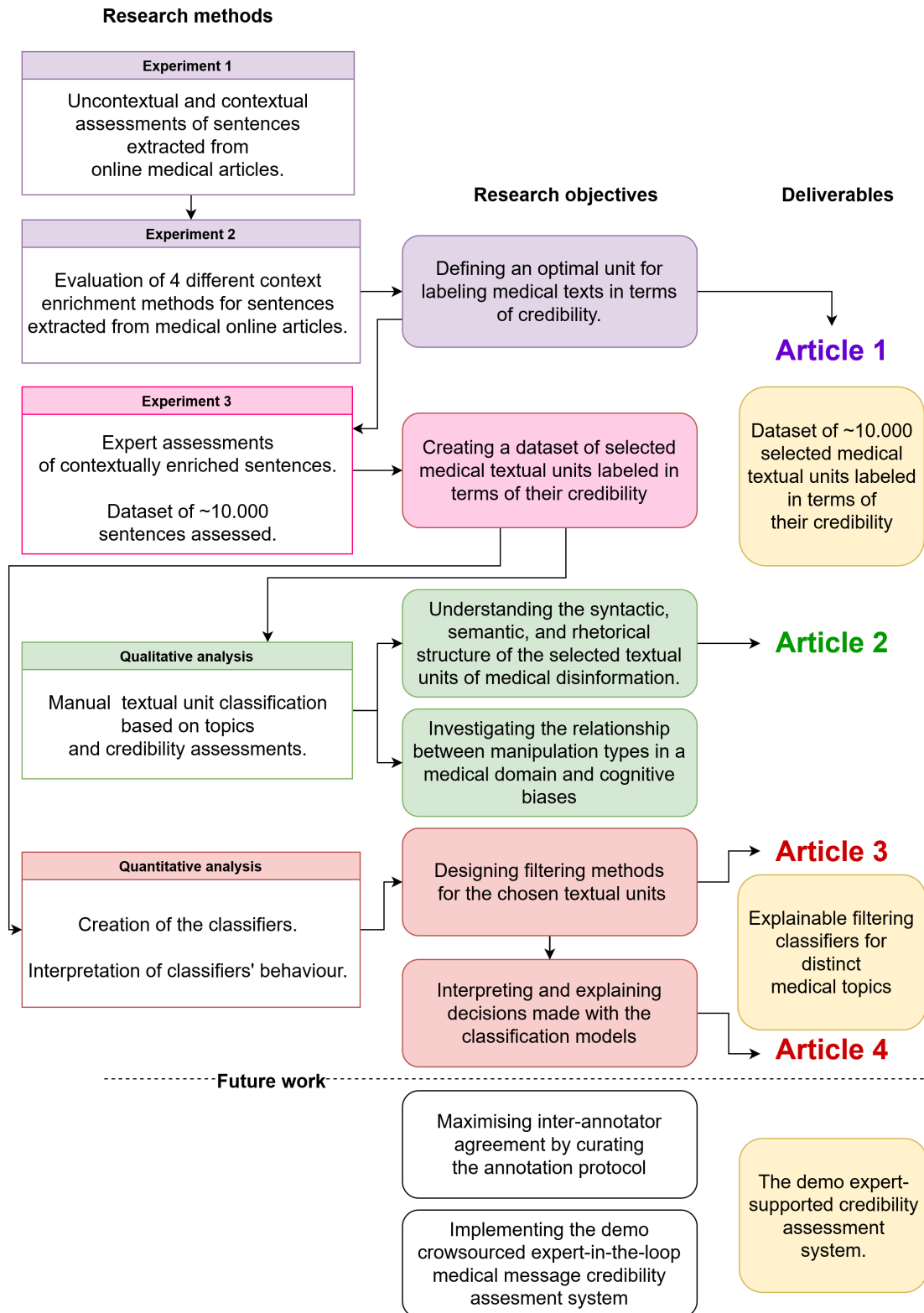


Figure 1.1: The workflow.

1.2.1 Defining an optimal unit for labeling online health information in terms of credibility.

When considering Webpage credibility assessment by a human expert, the most intuitive approach is to read the whole article and provide a single credibility label. While this may seem the most accurate procedure, it has some severe drawbacks. It requires the annotator sufficient time to read the full text and significant cognitive effort to process the credibility assessment of many sentences that interact with one another. As the expert's time is often limited, I decided to consider smaller chunks of text to be processed in a single annotation event for the efficient human-in-the-loop assessment system with high bandwidth. These shorter text units could be selected with a recommendation algorithm for expert annotation. A simple alternative to evaluating the whole article is to provide the expert with a single sentence at once. While this method has the advantage of speed, it has a drawback: the contextual information that may be indispensable for evaluation is missing. Therefore, the research objective was to find a compromise between a text that is too long and a text that is short enough to be labeled in terms of credibility without losing the necessary context.

1.2.2 Creating a dataset of selected medical textual units labeled in terms of their credibility

The dataset of around 10,000 annotations of statements related to selected medical subjects such as psychiatry or cardiology had to be constructed. Subject matter experts from corresponding medical fields had to be employed for annotation work. My goal was to release the labeled statement dataset to the general public. Furthermore, an active annotation framework for more efficient annotation of non-credible medical statements was meant to be the indirect result of the work on constructing the dataset.

1.2.3 Understanding the syntactic, semantic, and rhetorical structure of the new units of medical disinformation.

Extensive research has been done regarding features affecting the credibility of medical web articles available online. However, there is little research on specific manipulation techniques and rhetorical structures used when shorter text fragments (single sentences/statements) are considered. The present research aimed to fill this gap.

1.2.4 Investigating the relationship between manipulation types in a medical domain and cognitive biases

Identifying false content on the Web is a relatively new direction; thus the guidelines and ways of classifying dis- and misinformation vary from one research institution to another. Cognitive biases that increase susceptibility to this phenomenon are observed in both laymen and experts. A lack of knowledge about how experts react to cognitive biases and various types of manipulation in medical disinformation limits the possibility of creating a coherent annotation protocol and, thus, creating qualitative training datasets for machine learning.

The present research objective was to investigate to what extent medical misinformation differs from disinformation in the general news domain, highlighting the differences, finding similarities, and describing the aggregated classes using widespread psychological phenomena. It includes an overview of information science research into the types of disinformation, categories of manipulation, and persuasion.

1.2.5 Designing an annotation protocol for labeling medical content for a given textual unit

Automatic labeling of textual data by machine learning algorithms is only possible with a solid and carefully crafted training dataset. In constructing such a

dataset, subject matter experts should be involved and provided with a clear and precise annotation protocol. A suitable protocol should generate few inconsistencies between particular annotators. To obtain maximum inter-rater agreements, the annotators should be aware of their cognitive biases and the types of manipulation prevalent in medical misinformation. Results obtained while working on the previous research objective should be the basis for curating a better annotation protocol.

The work on this research objective is ongoing and thus is not a part of this dissertation. In Figure 1.1 it is moved to the Future Work section.

1.2.6 Designing filtering methods for selected textual units

Classification of medical messages as credible or non-credible is crucial for human life and well-being. Following medical advice contrary to the current medical guidelines can have severe and adverse effects. That is why the subject matter expert should decide on the credibility of a given message. However, a system with only humans involved in a classification task would have insufficient capacity for real-life scenarios. For this reason, information pre-filtering should be performed so that mostly unreliable statements should be chosen for the final stage of human annotation. The credibility classification should be performed following the criteria: precision for the positive (credible) class and the difference between the proportion of non-credible messages in the filtered dataset to the non-filtered (original) dataset (Negative predictive value versus the general negative samples proportion). Maximizing the first criterion ensures that the annotators skip little unreliable content. Maximizing the second criterion, on the other hand, significantly increases the capacity of the human-supported verification system.

1.2.7 Interpreting and explaining decisions made by classification models

From the usability perspective, the obtained classifiers should not only reach desirable performance characteristics but also justify their choices. Thus,

Chapter 1. Introduction

building explainable models that provide unambiguous feedback is one of the research objectives. It was essential to gain insights into a model's predictions at the semantic, syntactic, and language sentiment levels.

CHAPTER 2

Literature Review

One of the first projects in the field of computer science to deal with the problem of credibility was the Reconcile project carried out between 2014 and 2017 at the Polish-Japanese Academy of Information Technology and École Polytechnique Fédérale de Lausanne. The efforts of scientists working on the project are summarized in the book "Web Content Credibility" by the project manager, A. Wierzbicki [3]. However, it is still an immature field of research, particularly in terms of its practical applications such as accurate and scalable real-time disinformation recognition (see seminal review by Lazer et al. [4]).

Credibility is a term that encompasses many concepts. From fairly unambiguous values (solidity, reliability, honesty), to more ethereal ones, such as the quality or aesthetics of the content. In the monograph "Credibility in Information Retrieval", concerning the achievements in the field of credibility research until 2015 [5], as many as eight definitions of this concept were distinguished in various dictionaries of the English language. Seemingly similar definitions, sometimes they differ fundamentally. Some define credibility as a certain value (*"the value that makes people believe or trust someone"*), and some define credibility as a fact, a state of affairs (*"the fact that someone is trustworthy or entrusted to him"*).

Both of the above definitions refer to the publisher, which is the source of information. However, when facing the task of assessing the credibility of a text in isolation from its source, one should refer to a broader definition. Such a definition is given by the Merriam-Webster dictionary ¹. It states that credibility is *"the*

¹tinyurl.com/4f29w34h

Chapter 2. Literature Review

quality or power of inspiring belief". Such power may be a trait of a specific agent - an organization, a person, or a public institution. In the realms of the online world, it would therefore be related to concepts such as online trust, communities, persuasive design of a webpage, etc. On the other hand, such power may also be a feature of a single message and thus linked to language issues.

This idea was developed in the Hovland-Yale model, also known as the Yale attitude change model. It was created by psychologist Carl Hovland in the 1950s and describes three conditions under which people are most likely to change their attitudes in response to persuasive messages [6]. Those conditions include:

1. Source factors (e.g. trustworthiness, likeability, expertise, attractiveness).
2. Message factors (e.g. order of arguments, whether they are one or two-sided, perceived persuasiveness and/or intentions).
3. Audience factors (e.g. persuasiveness, intelligence, the self-esteem of the person).

I will further elaborate on the source and the message factors.

2.1 Source credibility

2.1.1 Source credibility in the general domain

From the source credibility perspective, some of the most important predictors of the perceived credibility of online content are trust and availability, especially in the context of medical information portals [7].

The question "*What is online trust?*" is formulated in the works of Corritore et. al. [8] as well as Artz and Gil [9]. In their work summarizing the research on trust, Artz and Gil propose the division of the so-called trust in information sources into:

- trust issues on the Web

Chapter 2. Literature Review

- trust issues on the Semantic Web
- network-related trust (web of links on websites)
- information filtering for trust
- semantic Web filtering
- subjectivity analysis
- origin of information
- content trust
- webpage design and human aspects.

This dissertation focuses primarily on the aspects concerning **information filtering, subjectivity analysis, and content trust**.

Information filtering is related to the concept of quality. Providing high-quality information seems to be a common goal of researchers making efforts to filter out relevant content for Web users. It is also strongly correlated with trust. [10] describes this issue in detail, emphasizing that enormous amounts of content on the Web quickly become obsolete.

Subjectivity analysis is a broad topic that is currently dealt with mainly in the field of natural language processing (NLP) research. Because it is an aspect related to both the source credibility and message credibility, it will be discussed in more detail later in this chapter while discussing language modeling.

Content trust decision-making, as described by Castelfranchi et al. [11] is a complex process that includes four aspects:

- direct experience,
- categorization (generalization about or from something known),
- reasoning (applying common sense or rules to verify the truth)
- reputation.

It can be seen that elements of online trust are more related to the characteristics of the recipient than to the rated entity. Such characteristics, e.g. the

Chapter 2. Literature Review

socio-economic situation or various psychological factors, are described in [12] and [13].

There are also studies showing that people accept social media as a trustworthy source of expert information [14] and that a generation of people used to deal with highly interactive websites generally perceive less credible content as more credible [15].

Social media also share a phenomenon known as "*the propagation of trust and distrust*" [16]. Researchers at the University of California confirmed that most of the user feedback on trustworthiness is based on feedback from others [17], which was modeled in [18] using the trust propagation theory. Taking this into account, researchers take advantage of a wide range of possibilities when it comes to creating assessment tools based on the analysis of social networks. For example, an effective ranking tool for assessing the credibility of blogs was created in [19], and a tool for assessing Twitter entries [20].

2.1.2 Source credibility in the medical domain

This dissertation presents efforts to automate the assessment of the credibility of online health information. Therefore, it is necessary to emphasize the difference between the assessment of the credibility of this type of content and texts from the general news domain. Stating the truth in the medical field is generally more difficult because it is burdened with greater uncertainty, conditionality, and temporariness (medical guidelines can change dramatically from month to month). Thus, more emphasis in assessing the quality of medical information is placed on the accuracy of the reporting, the correctness of the cited statistics, and the completeness of the information. Quality assessment coding schemes for lay medical articles were already proposed in the 90s under the DISCERN project [21], and as the Health on the Net principles. The Health On the Net Foundation (HON) "was created in May 1996, during the beginning of the World Wide Web, from a collective decision by health specialists"[22] and remains the only tool to focus solely on the information source (e.g. the Web portal with multiple articles). HON Principles include:

Chapter 2. Literature Review

- Principle 1: **Authority** - Give qualifications of authors;
- Principle 2: **Complementarity** - Information to support, not replace;
- Principle 3: **Confidentiality** - Respect the privacy of site users;
- Principle 4: **Attribution** - Cite the sources and dates of medical information;
- Principle 5: **Justifiability** - Justification of claims / balanced and objective claims;
- Principle 6: **Transparency** - Accessibility, provide valid contact details;
- Principle 7: **Financial disclosure** - Provide details of funding;
- Principle 8: **Advertising** - Clearly distinguish advertising from editorial content. Based on the principles given source can obtain a time-limited certificate of quality.

Other guidelines incorporate several source-level criteria, but refer mostly to the written content or the presentation aspects (e.g. presence of the charts and pictures). The comparison of those criteria will be listed in 2.2.2.

Many works derive source-level credibility metrics to further assess message credibility or mix source-level features with message-level features as inputs for classifiers. Thus, the review of works that automate the assessment process is also part of chapter 2.2.2.

2.2 Message credibility

The concept of *message credibility* is intuitively linked to the notion of truth. However, none of the available credibility definitions unequivocally indicates truthfulness as a necessary element of credible information. This implies a lot of ambiguity in the world of science because the task of assessing information credibility is often confused with a separate task of detecting the so-called *fake news*. Meanwhile, A. Wierzbicki gives the following definition: *Credibility can be defined as a signal that is received by any recipient of the information and may be used by that recipient to decide whether to accept or reject the information* [3]. This means that whenever a Web user receives information (potentially false), they make

Chapter 2. Literature Review

an internal credibility assessment. Factors influencing the shaping of this signal depend on several characteristics, which are described later in this text. In the field of research on Web credibility, some works distinguish truthfulness as a separate characteristic. Sometimes it is treated only as an element of credibility. While checking the truthfulness of information, the most problematic aspect is whether we are able to verify the facts contained in a given statement. **Intentionally false information posted on the Web is often written in such a way that its verification is difficult or even impossible.** Credibility, on the other hand, is something that we can assess regardless of the message's verifiability.

Let me now review the state-of-the-art methods developed to automate the credibility of messages in the general news domain, to later move to the less broad topic of the medical domain.

2.2.1 Message credibility assessment support tools

When developing tools to help detect unreliable content, researchers use the entire conceptual framework described in previous sections. There are separate systems to detect individual features that are components of credibility assessment, as well as hybrid technologies that directly address the main issue. Researchers approach solving the problem from different angles, focusing on the network of connections between the recipients of the content, the usefulness of the websites that convey the content, and the text layer itself. One of the most recent systematic literature reviews of this matter, collectively dealing with the issue of veracity assessment [23] groups research directions into:

- utilizing implicit features,
- employing explicit fact-checking,
- the appeal to authority method.

[23] claim that the *implicit features* approach is by far the most common. The idea behind is that claims that are non-credible differ from claims that are credible in some non-veracity properties.

Chapter 2. Literature Review

On the other hand, in the *explicit fact-checking approach* the idea is to compare a claim with some existing body of knowledge so as to determine if it is veridical. Lastly, in the rarest direction called *the appeal to authority approach*, the idea is that a claim is veridical if it is claimed by an authoritative source. For example, a photo can be trusted if shared by a trusted source 30 min after the event [24], and a claim can be considered veridical if supported by the majority [25] or by verified news channels [26].

Below, the *implicit features* and *explicit fact-checking approach* will be discussed in more detail.

Explicit fact checking

Despite difficulties with the automatic verification of information truthfulness, there are many papers describing efforts of scientists to achieve this task, defined as "automated fact-checking". It focuses mainly on work related to the fight against fake news. This task is possible to carry out when we extract from the examined text a sentence that contains just one statement about a specific fact. This statement in turn must be formulated in such a way that it can be assigned a binary label - true or false. Tagging compound sentences and opinions must be treated as a separate problem [27]. Fortunately, it is possible to extract simple statements automatically [28]. When we have already singled out simple statements that meet the above assumptions, checking their truthfulness usually takes place by comparing the representation of the facts to an external knowledge base. Domain knowledge bases are very useful in this task. However, it is also possible to use external databases, even the unstructured ones [29].

Some approaches to automatic fact-checking rely on statistical methods [30], while others rely on comparisons with graph representations of statements and therefore require graph knowledge bases [29], [31]. The second type of approach provides result of as good a quality as that of a continuously maintained, adequately broad knowledge base. For this reason, many researchers use the resources of DBpedia - a graph knowledge base that gathers encyclopedic knowledge based on the resources of Wikipedia [31], [32], [33], [34], [35]. Unfortunately, it has one

Chapter 2. Literature Review

serious drawback - it contains no negative examples. This means that in the first place, it should be assumed that the lack of a statement in DBpedia immediately implies its falsehood, which is not necessarily true.

Works presenting methods that support automatic fact-checking for structured data include:

- the already mentioned description of the graph knowledge base constructed using the verified statements [32],[36]
- a system that uses KnowLife knowledge graph (<http://knowlife.mpi-inf.mpg.de/>) together with text embeddings to feed the neural network for website classification [37]
- a semi-automatic rumor verification system in which journalists play an important role (as experts involved in the final labeling) [38],
- a system that extracts facts, statements about those facts, and the beliefs of users from discussion forums (BeLink [39]). It is based on the W3C RDF and Linked Open Data standards and enables the construction of queries in SPARQL. The system allows for tracking the propagation of statements in the network.
- an ontology-based tool for detecting anti-vaccine claims throughout the Web [40].

Publicly available fake news datasets include:

- MultiFC - a collection of statements and appropriately tagged sources, on the basis of which one can prove the truthfulness or falsehood of a given statement [41]
- "Liar, liar pants on fire" - a decade-long, 12.8K manually labeled short statements in various contexts which provides detailed analysis report and links to source documents for each case [42].
- A manually annotated dataset of 10,700 social media posts and articles of real and fake news on COVID-19. Fake claims are collected from Politifact, NewsChecker, Boomlive, etc., and from tools like Google fact-check-explorer,

Chapter 2. Literature Review

and IFCN chatbot. The real news is collected from Twitter using verified Twitter handles.[43]

- Datasets from the Fact Extraction and VERification shared task (FEVER) [44]. According to the Authors, the 2018 version "consists of 185,445 claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. The claims are classified as Supported, Refuted or NotEnoughInfo. For the first two classes, the annotators also recorded the sentence(s) forming the necessary evidence for their judgment."
- *CheckThat!: Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News* provides a dataset of news articles scraped from various fact-checking websites with merged "True", "Partially true", "False" or "other" labels [45]. It is a part of the Conference and Labs of the Evaluation Forum (CLEF) initiative.

As it has already been stated, the key drawback of explicit fact-checking is that the majority of mis- and disinformation volume is unverifiable. Secondly, it is not always possible to retrieve simple claims from unstructured text. Thirdly, maintaining structured Knowledge Bases of verified claims is costly, and even in a perfect scenario of a Base that contains every possible claim stated in the World Wide Web quick detection of misinformation in the new emergent topic would be hard to accomplish.

Implicit features method

Examples of tools for detecting phenomena related to credibility are models for detecting controversy in the text [46]. Automatic support for manual credibility assessments, on the other hand, may be based on detecting relevant sentences those worth verifying and containing statements that can be proved to be true. Such an attempt was carried out in the [47] study, which proposed the use of the TextRank algorithm to extract essential sentences. It was shown that the assessment of reliability and significance are statistically related. Extracting check-worthy claims is also one of the tasks in the CheckThat! Lab of the aforementioned CLEF

Chapter 2. Literature Review

initiative [48].

Excellent example of message credibility assessment is studies related to the credibility of information on blogs. An up-to-date literature review of this matter is given by Wagle et. al. in their work describing an explainable AI system based on a case study of online misinformation on beauty health [49]. They mention extensive research that has been done on the content of web blogs to determine credibility based on different factors. For example, credibility signals can be identified in the author's sentiments, the expertise reflected in the content, as well as the readability, grammar, and vocabulary used. It has been observed that blogs presenting reliable information contain more words, sentences, and numbers than disinformation blogs [50]. Linear Regression and Neural Networks are the two approaches suggested for web page credibility by Jaworski in 2014 [51]. Manjula and M. S. Vijaya suggested a predictive model based on deep neural networks using an elaborate dataset of health pages [52]. Content factors like readability, freshness, and duplicates were extracted from the text body. Then they were redefined, and a new set of features was self-learned through the deep layers of the neural network. As to the more recent examples, dominant approaches in the "CheckThat! CLEF shared task for fake news detection" were transformer architectures[53, 54, 55]. Mainly Global Vectors (Glove) [56], Bidirectional Encoder Representations from Transformers (BERT)[57] and optimized BERT ("Robust BERT" called RoBERTa) [58]. There were also experiments carried out with traditional text processing methods such as TFIDF, for example in a combination with Naive Bayes [53]. Surprisingly, there have been no attempts to model knowledge with the so-called semantic technology (e.g. transforming a text into triplets) or to extract statements that could be checked against knowledge bases.

2.2.2 Message credibility assessment in the medical domain

Guidelines to assess health-related written online content have to comply with the rapidly-evolving online reality, thus new tools and updates are designed every few years. The already discussed DISCERN and HoN projects emerged in the 90s, but newer tools are also available. For example, the Ensuring Quality Information for Patients [59] (EQIP, 2004) and Evidence-Based Patient Information

Chapter 2. Literature Review

(EBPI, 2010) [60], or Good practice guidelines for health information (GPPI, 2016) [61], to name a few. Keselman et al. [62] propose different credibility assessment criteria based on 25 online articles regarding Type 2 diabetes. Those criteria (objectivity, emotional appeal, promises, and certainty) can be automatically captured by language models and lexicon-based machine learning.

Each of the tools listed represents slightly different objectives and often radically different labeling strategies:

- DISCERN aims at assessing the **quality** of written information on treatment choices for a health problem. The Authors of [63] developed an explicit scheme for developing a 5-star quality rating system for consumer health information based on DISCERN. The helper questions for Discern are created based on the assumption that a good quality publication about treatment choices will: (a) Have explicit aims (b) Achieve its aims (c) Be relevant to consumers (d) Make sources of information explicit (e) Make the date of information explicit (f) Be balanced and unbiased (g) List additional sources of information (h) Refer to areas of uncertainty (i) Describe how treatment works (j) Describe the benefits of treatment (k) Describe the risks of treatment (l) Describe what would happen without treatment (m) Describe the effects of treatment choices on the overall quality of life (n) Make it clear there may be more than one possible treatment choice, and (o) Provide support for shared decision-making.
- EQUIP tool, on the other hand, was developed to assess the **presentation quality** of all types of written health care information. Its criteria state that: (a) the information should be clearly communicated; (b) be evidence-based; and (c) involve patients in the development of the materials.
- EBPI also focuses on **presentation quality** aspects and scores articles based on: 1. Content of information and meta-information, 2. Quality of the evidence, 3. Patient-oriented outcome measures, 4. Presentation of numerical data, 5. Verbal presentation of risks, 6. Diagrams, graphics, and charts, 7. Loss- and gain-framing, 8. Pictures and drawings, 9. Patient narratives, 10. Cultural aspects, 11. Layout, 12. Language, 13. Development process.
- GPPI puts forward methodological aspects to be considered when developing health information, it is therefore questionable whether it fits the definition

Chapter 2. Literature Review

of an assessment tool for the existing written materials.

Surprisingly, the tools used by medical journalists and practitioners are not very common within the computer science society. For example, the annotation protocol for the TREC health Misinformation shared task [64] included their own annotation protocol for the credibility labeling of the documents claimed as useful by annotators [65].

Professional tools are not widely used, but they are not nonexistent in computer science studies. The examples include:

- AutoDiscern project that aimed at automating assessment across 6 DISCERN criteria [66],
- the work of Shah *et. al.* which incorporates several assessment tools to extract 7 criteria for which classification models are created and tested among vaccine-related webpages [67], and
- Sondhi *et. al.* who developed a gold standard dataset using the standard reliability criteria defined by the Health on Net Foundation to achieve 80% accuracy in automatically predicting the reliability of medical webpages.

There is also the set of principles developed by the health journalist Gary Schwitzer under the project **HealthNewsReview.org**[68]. It was one of a "web-based projects that rated the completeness, accuracy, and balance of news stories that included claims about medical treatments, tests, products, and procedures". The rating instrument was rather complex and included ten criteria used by Australian and Canadian Media Doctor sites:

1. Adequately discusses costs.
2. Quantifies benefits.
3. Adequately explains and quantifies potential harms.
4. Compares the new idea with existing alternatives.
5. Seeks out independent sources and discloses potential conflicts of interest.
6. Avoids disease mongering.

Chapter 2. Literature Review

7. Reviews the study methodology or the quality of the evidence.
8. Establishes the true novelty of the idea
9. Establishes the availability of the product or procedure.
10. Appears not to rely solely or largely on a news release.

Although due to the lack of funding, the project ceased to publish new reviews in 2018, a dataset of around 2000 health-related news stories and their ratings is still available. There are several studies newer than 2018 that use it:

- as the main dataset for training machine learning quality classifiers [69, 70],
or
- as a supporting dataset for training the credible health-related articles search engine component [71]

In the works that do not take advantage of any professional credibility annotation tools for creating datasets, a common practice is to derive message credibility labels from source credibility. An example is the widely known CoAID dataset that contains 4251 news stories and claims about the COVID-19 pandemic [72]. Reliable texts are collected from 9 reliable media outlets, and unreliable ones are collected from Websites marked as non-credible by several fact-checking organizations. As for the COVID-related misinformation datasets - there are several studies that attempt automatic misinformation detection of claims using their dedicated datasets with crowd-sourced labels. For example, CoVerifysystem [73] uses a dataset of Twitter posts with the majority vote credibility label.

The topic of health-related message credibility assessments also includes some works for narrow domains, such as:

- [74] the Support Vector Machines algorithm with a bag of words representation achieving 93% efficiency in recognizing content with harmful information on alternative cancer treatments.
- [75] a combination of a bag of words and statements about depression used to calculate significance coefficients for each article to assess the reliability of each observation using ranking methods.

Chapter 2. Literature Review

As to other works, the approaches to automatic classification of online medical misinformation differ depending on the medium and the content type. Most studies employ content analysis, social network analysis, or experiments drawing from disciplinary paradigms [76]. Thus, it is not easy to distinguish whether the source or the message credibility is taken into account. Listed below are the works which incorporate lexical and semantic features (attributed mostly to message assessments) together with network-related features (attributed mostly to source assessments and taking into account both social media and the Web).

Zhao *et. al.* use so-called peripheral-level features to classify medical misinformation [77]. They include linguistic features (the length of a post, presence of a picture, the inclusion of an URL, content similarity with the main discussion thread), sentiment features (both corpus-based and language model-based), and behavioral features (discussion initiation, interaction engagement, influential scope). Peripheral-level features proved to be useful for detecting the spread of false medical information during the Zika virus epidemic [78]. Stylistic features can be used to identify hoaxes presented as genuine news articles and promoted on social media [79]. Along with identifying hoaxes, it is possible to single out social media users who are prone to disseminating these hoaxes among their peers [80]. An ensemble of word sentiment features and online popularity metrics is successfully applied to distinguish between online anti- and pro-vaccine article headlines [81].

An applied machine learning-based approach, called *MedFact*, is proposed in [82], where the authors present an algorithm for trustworthy medical information recommendation. The *MedFact* algorithm relies on keyword extraction techniques to assess the factual accuracy of statements posted in online health-related forums.

The team from the Max Planck Institute successfully approached the identification of the credibility of the content posted by users of online forums on drug use [83]. The results turned out to be so effective that on the basis of the extracted information it was possible to conclude about the side effects of using a combination of different drugs before clinical trials were launched.

The ensemble model for medical message credibility assessment is applied for Twitter data in [84]. First, tweets are retrieved by the keyword-based extraction method linked to Wikipedia API. Then, the process of automatic verification is

Chapter 2. Literature Review

performed step-wise. First, the arbitrarily stated credibility of the source is taken into account. Non-verified sources, on the other hand, are compared to trusted tweets by context and sentiment and if the tweets are dissimilar then they go through a machine learning classifier that analyzes the user-based, content-based, and network-based features. Liu et. al. [85] classify popular medical content in Chinese as deceptive or not. Features based on text analysis together with those related to source analysis are used in the classification. Here, even though the Authors present mainly the approach of *message credibility* modeling, they arbitrarily assign ground truth labels to their dataset based on the source of the article.

In [86] The Authors undertake the task of classifying health-related press releases. Working on a collection of articles from reliable and unreliable sources, The Authors distinguished structural, semantic, and thematic features, and then built a highly effective classifier based on them (F1 measure 96%). The features that power the model are:

- identification of distinctive "click-bait" titles;
- the subject matter (topics where the articles are more likely to be unreliable include, for example, nutrition, slimming, or caring for the skin);
- the content of linked research.

In [87] The Author presents a model that classifies information in terms of whether it contains meaningful statements from the perspective of the particular domain knowledge. It should be noted that the relevance of information is also a component of credibility.

More advanced methods of online medical information evaluation include video analysis (extracting medical knowledge from YouTube videos [88]), detecting misinformation based on multi-modal features (both text and graphics [89]), and website topic classification. The last approach was successfully applied by [90], [69] using the topic analysis (either Latent Dirichlet Annotation or Term-Frequency). In addition, Afsana *et al.* use linguistic features, such as word counts, named entities, semantic coherence of articles, the Linguistic Inquiry Word Count (LIWC), and external metrics such as citation counts and Web ranking of a document. [91]

Chapter 2. Literature Review

incorporates an even richer set of features for article-level credibility assessment. The Authors take into account: URLs, titles, keywords, text, images, tags, authors, date news reviews rating, the ground truth of rating criteria, explanations of the ground truth, category, summary, descriptions, source, social engagements, tweets about the news source, as well as the tweet's replies, retweets, user network, profiles, timelines, followings, and followers.

2.2.3 Language modeling as a support tool for message credibility assessment

As it has already been stated, message credibility is the focal point of this dissertation. Content-based assessments play a vital role in this process, so the selection of the proper tools was essential. Here I present a review of the state-of-the-art language modeling tools to support credibility assessments of online health information.

Language modeling, especially regarding the models that incorporate modern deep learning architectures, is a topic that has gained a lot of attention recently. As it has been said, credibility analysis of textual content relies on the implicit features related to sentiment, propaganda, persuasive language, and others. Moreover, as deceptive and/or obsolete information tends to be repeated throughout the Web, calculating semantic similarities is also a practical issue in detecting non-credible content. We can view those tasks as subtasks for implicit veracity assessment. Deep learning language models are widely used as part of each task. Therefore, I provide a brief introduction to the subject - starting with basic word vectorization techniques and finishing with the more sophisticated sentence representations that were used in works that form part of this dissertation.

Since the seminal work of Mikolov et al. [92], deep learning-based word embeddings have revolutionized the space of natural language processing. After the initial success of the word2vec algorithm, numerous alternatives have been introduced: GloVe embeddings trained via matrix factorization [56], embeddings trained on sentence dependency parse trees [93], embeddings in the hyperbolic space [94], sub-word embeddings [95], and many more. A common feature of these

Chapter 2. Literature Review

embeddings is the static assignment of dense vector representations to words. Each word receives the same embedding vector irrespective of the context in which the word appears in a sentence. These static embeddings can be used to create representations for larger text units, such as sentences, paragraphs, or documents. However, static embeddings are inherently unable to capture the intricacies hidden in the structure of the language and encoded in the context in which each word appears. Consider two sentences: "A photo reveals significant damage to the tissue" and "Please do not throw used tissues into the toilet". The word "tissue" will receive the same vector even though the context allows for disambiguating the meaning. To mitigate this limitation, modern language models depend on deep neural network architectures to calculate accurate, context-dependent word and sentence embeddings. First, context-dependent language models utilized either the long short-term memory (LSTM) network architecture [96] or gated recurrent unit (GRU) networks [97] to capture contextual dependencies between words appearing in a sentence. In other words, contrary to static word embeddings, context-dependent language models calculate an embedding word vector based on the context (i.e. words surrounding the embedded words). In the aforementioned example, the word "tissue" would receive two different vector representations: in the first sentence, the vector for the word "tissue" would be much closer to vectors of words such as "skin" or "cell", whereas in the second sentence the vector for the word "tissue" would be closer to the vector of the word "handkerchief". These early recurrent architectures, however, suffered from performance drawbacks, and in 2018 they have been replaced by the transformer architecture [98]. This architecture allowed for training much better embeddings, such as Google's Universal Sentence Encoder [99] or the (infamous) GPT-3 [100]. The current state-of-the-art language model, called BERT [57] (Bidirectional Encoder Representations from Transformers), produces continuous word vector representations by training the neural network using two parallel objectives: guessing the masked word in a sentence (i.e. trying to predict the word based on the context), and deciding whether two sentences appear one after another. Given such training objectives, the network applies similar weights to the nodes regarding input words that appear in a similar context. Sentence-BERT (sBERT) [101] is a straightforward extension of the original BERT architecture for creating sentence embeddings. This model is based on Siamese BERT networks [102] (two identical models trained simultaneously) that are fine-

tuned to the Natural Language Inference and the Semantic Textual Similarity tasks. The model serves as an encoder for sentences, which calculates vector representations of sentences so that semantically similar sentences have low cosine distance in the latent embedding space. This is both more efficient and produces semantically richer sentence representations than simply averaging the vectors of words that appear in each sentence.

2.2.4 Context enrichment for assessing short message credibility

Modeling of the context is present in many solutions for downstream natural language processing tasks. However, in most cases, it serves only as an intermediate tool. It enhances the performance of different solutions aimed at, for example, semantic text similarity prediction, sentiment analysis, and machine translation. The context is usually coded in such a way that is not possible to interpret by a human but only by a neural network that processes this context. In [103] surrounding sentences are used to better learn vector representations of the input sentence, similarly to the way in which the word2vec algorithm learns the representation of the word. In [104], on the other hand, context summarized in a hierarchical way is integrated with the neural machine translation model as a source for updating decoder states. In [105] the authors take advantage of contextual relations among sentences so as to improve the performance of sentence regression for text summarization.

In the studies that form part of this dissertation, the aim was to retrieve the context directly, so that it could be later accessed in a human-readable format. A variety of methods exist that include direct extraction of the context. The cloze-style reading comprehension problem has recently become a well-known baseline NLP task, where the level of text understanding by the system is tested by asking it questions, the answers to which can be inferred from the document. In [106] the query is designed in a form of a short sentence that summarizes a statement that appears in the text but lacks one named entity. Predicting the missing component requires a deep understanding of the context. The Authors take advantage of the popular deep-learning architectures with recurrent neural networks and pay close

Chapter 2. Literature Review

attention to solving this problem. Their approach is to review and simplify the existing solutions, such as Pointer Networks [107] or Memory Neural Networks for text comprehension [108], which has resulted in a new state-of-the-art text comprehension algorithm.

Aside from the aforementioned NLP methods, there is a whole other branch of methods that utilize rule-based algorithms for context extraction. These methods are used to support decision-making by retrieving the context from electronic medical reports. For example, the ConText algorithm [109] derives information such as negation, experiencer, and temporality of the medical condition. One of the methods presented in this dissertation is also rule-based but as the addition to the more general keyword-based approach.

Some of the previous works, designed to support the credibility assessment of query, take advantage of the automatically retrieved context. [110] uses global context (derived from the whole set of documents retrieved by the search engine) to prompt the user with sentences that may indicate controversy related to the given query, whereas [111] uses context to provide the information whether given article supports or rejects the statement contained in the query. Unlike my approach, both studies are focused on a regular Web user (not an expert) and treat the query as a whole, not as part of a larger content.

2.3 Implications

According to [112], credible websites can promote unreliable content and *vice versa*. Therefore, I decided to focus on message credibility rather than source credibility. Moreover, considering the short lifespan of medical advice given the rapidly changing guidelines, long health texts can be composed of reliable and unreliable statements. For this reason, I decided to search for a new, short unit for message credibility assessment. It is a new approach compared to previous work. Other research concerns datasets of whole news stories, webpages, or social media entries, which use the available information units without considering their usefulness for credibility evaluation by medical experts. Compared to the existing work, my approach also differs in obtaining labels of ground truth credibility. While

Chapter 2. Literature Review

in other approaches, data is labeled by laypersons in a majority vote manner or derived from the source credibility labels, I decided to focus on the credibility assessments of short medical messages provided by experts. This approach makes it possible to gain valuable insights into the rhetorical structure of the message. Moreover, it should be noted that most existing tools for assessing medical written content are focused on overall quality rather than credibility. While the quality is undoubtedly related to credibility, in my research, expert practitioners were able to focus on credibility evaluations using evidence-based medicine. The history of the HealthNewsReview.org initiative shows that reviewing whole news stories is resource-hungry and offers low efficiency. For a potential crowd-sourced expert-in-the-loop annotation system, using shorter message text units instead of full-length articles would substantially increase the capacity of the system. For a medical expert with minimal time, providing a credibility label for a short message is more effortless than assessing the whole document. Additionally, in an effective alert system for unreliable medical content, the end user should be provided with explanations of the algorithmic choices. For long medical texts, such explanations would be blurred, as credible and non-credible statements can be mixed and interact with one another. Finally, the review of related work points out that the human-in-the-loop rating system for online health content has yet to be considered. Most of the related work focuses on full automation, which, in my opinion, is too dangerous because of the potentially high cost of errors. Instead, I focused on partially automating the task: designing performance measures, text units, and exploratory power to make the system usable for the end user and the expert annotator.

To sum up, I conclude from the literature review that

- designing an annotation protocol and support system **aimed at expert annotators**,
- curating a credibility large dataset of **short information units** with medical expert credibility evaluations,
- and designing credibility classification algorithms for such shorter units of information

Chapter 2. Literature Review

are the issues the research community has not yet tackled.

CHAPTER 3

Contributions

Pursuing the research objectives stated in Section 1.2, I conducted several experiments to investigate the credibility evaluation of different text units and to create datasets. Following the experimental results analysis, I constructed the annotation protocol and optimized the annotation flow for subject matter experts. I prepared an open-sourced high-quality dataset of medical messages labeled as credible, non-credible, or neutral. Finally, I created a set of classifiers that increase the capacity of the expert-in-the-loop system for verification of medical online content credibility.

The published articles describing these contributions in detail [113, 114, 115, 116] are part of this dissertation (see section 5).

I shall describe the research contributions of this thesis by referring to the research objectives.

3.1 Contribution 1. Proposing sentence triplets as a unit of medical credibility evaluation.

This contribution relates to Objective 1: **Defining an optimal unit for labeling online health information in terms of its credibility.**

I checked that the Context Window approach performs best in terms of a trade-off between too short a message (with the missing context) and too long a message (with a long processing time for the annotator). The Context Window

Chapter 3. Contributions

method is described in the article "**Enriching the Context: Methods of Improving the Non-contextual Assessment of Sentence Credibility**". The article was presented during the **International Conference on Web Information Systems Engineering 2019 (Core A)** and published in "Lecture Notes in Computer Science" (LNCS, volume 11881).

After a qualitative analysis of sentences marked by experts as impossible to assess without context in a preliminary study, I designed and evaluated four context-filling methods corresponding to identified types of missing contexts.

Subject matter experts were asked to rate the credibility of the sentences that were previously found impossible to evaluate without the context using each method separately. They were confronted with the sentences with added context summaries to compare how the automatically extracted context changes their perception of each sentence. Those methods include:

1. **Context window (CW)** Two preceding sentences and one following sentence formed the context summary in this method.
2. **TF-IDF Keywords + rule-based method of supplementing the meaning of pronouns** Term Frequency-Inverse Document Frequency (TF-IDF) statistic was used to retrieve 5 most relevant words from each article. It was then attached to the sentence as a keyword set. Then, for all sentences that contained pronouns, the rule-based method of supplementing the meaning of pronouns was applied. The algorithm for the rule-based method is described in the following article [113].
3. **TextRank Keywords + rule-based method of supplementing the meaning of pronouns** Instead of calculating the Tf-Idf scores for words from the entire document, only the 3 most relevant sentences from the entire article were used. This method was applied to focus on the most relevant parts of a document. This was important because some documents may be excessively long. These three sentences with the highest TextRank score were selected based on the algorithm as described by [117]. Next, the same rule-based method was used to complement the meaning of pronouns.
4. **Coreference resolution** Coreference resolution aims to identify words or

groups of words linked to the same concept. In order to perform coreference resolution, I used the Multiservice [118] web service developed for the Polish language by researchers from the Clarin project (<https://clarin-pl.eu>). All pronouns found in sentences were supplemented with the corresponding entities indicated by the algorithm.

Evaluating the performance metrics for the methods mentioned above and a verdict about the optimal method and a resulting sentence presentation forms part of the article above. The CW method was used to form the unit of information used in later experiments. I will later refer to such units as **sentence triplets**.

3.2 Contribution 2. Creation of a dataset of 10,000 sentences from online medical articles with credibility evaluations made by medical experts using a new annotation protocol

This contribution is related to Objectives 2, 3, and 5:

- **Objective 2: Creating a dataset of selected medical textual units labeled in terms of their credibility**
- **Objective 3: Understanding the syntactic, semantic, and rhetorical structure of the new units of medical disinformation**
- **Objective 5: Designing an annotation protocol for labeling medical content for a given textual unit**

Objectives 2 and 3 were achieved and described in the publication **Active Annotation in Evaluating the Credibility of Web-Based Medical Information: Guidelines for Creating Training Data Sets for Machine Learning** published in the **Journal of Medical Internet Research, Medical Informatics (IF 3.23)** [114]. The original contributions presented in the paper mentioned above are as follows:

Chapter 3. Contributions

1. An annotation schema, an annotation protocol, and a unique annotated dataset comprising 10,000 sentences taken from web-based content on medical issues labeled by medical experts as credible, non-credible, or neutral. The entire dataset is available in a public repository ¹.
2. A method for ranking sentences submitted to medical experts for labeling. Our active annotation method increases the likelihood that medical experts will identify non-credible sentences and thus optimizes the use of medical experts' time.
3. A qualitative analysis of the labeled dataset. I discovered four distinct narratives (syntactic and semantic) in the non-credible statements. These narratives can be further used to discern non-credible statements in medicine other than the areas covered by our dataset.

To construct the dataset, medical experts evaluated the credibility of sentences with the following set of labels and the corresponding instruction:

- CRED (credible): the sentence is reliable; does not raise major objections; it contains verifiable information from the medical domain
- NONCRED (not credible): the sentence contains false or unverifiable information; it contains persuasion contrary to current medical recommendations; it contains outdated information
- NEU (neutral): the sentence does not contain factual information (eg, it is a question); it is not related to medicine

Publication [114] includes dataset statistics, detailed information about record metadata, and a detailed annotation protocol (e.g. examples, helper questions, and additional tags for non-credible sentences).

¹https://github.com/alenabozny/medical_credibility_corpus

3.3 Contribution 3. Identification and description of manipulation techniques prevalent in online health information content

This contribution relates to Objective 3: **Understand the rhetorical, syntactic, and semantic structure of unreliable medical health content.**

The manipulation techniques identified after a qualitative analysis of the "Statins & cholesterol" topic include:

1. Slippery slope: The sentence is factually true, but the consequences of the presented fact are exaggerated.
2. Hedging: The sentence is factually incorrect, but a part of it softens the overtone of the presented statement.
3. Suggested negative consequences: The sentence is mostly factually accurate, but given the context of the expert's experience, there is a risk that the presented information may lead the patient to act contrary to current medical guidelines.
4. Twisting words: the presence of a single word changes the overtone of the sentence.

For further elaboration and examples, please refer to [114].

Aiming to create a universal classification system, I attempted to classify medical misinformation by referring to the existing annotation protocols for disinformation in the general news domain. I investigated to what extent medical misinformation is different.

I considered the disinformation classifications proposed by:

1. **DebunkEU** (debunkeu.org) - "an independent technological analytical center and an NGO, whose main task is to research disinformation in the public space and execute educational media literacy campaigns"[119]; Debunk is supported by DELFI, an information portal operating in the Baltic States, which was

Chapter 3. Contributions

founded in 1999 and in 2007 was acquired by the Estonian Ekspress Group; Debunk is also supported by Digital News Initiative, which is a European organization created by Google to "support high-quality journalism through technology and innovation" ("Digital News Initiative: €20 million of funding for innovation in the news". Google. 2017-12-13. Retrieved 2018-07-10.)

2. **EUvsDisinfo** is the flagship project of the European External Action Service's East StratCom Task Force([opens in a new tab](#)). It was established in 2015 to better forecast, address, and respond to the Russian Federation's ongoing disinformation campaigns affecting the European Union, its Member States, and countries in the shared neighborhood. (cite EuvsDisinfo page)
3. My own medical disinformation classification as proposed in [114].

- **Ad. 1.** DebunkEU classification can be found in [119] and consists of:
 1. **HYPERBOLIZATION** - The information is exaggerated or presented from two perspectives, purposely polarizing the audience. Arguments are based only on specific assumptions that are partially true, with one side having a lot of negative/positive connotations to discredit the opposite view.
 2. **SELECTION** - Information presented out of context selectively and intentionally leaves out important aspects of the situation.
 3. **FORGERY** - Information presented with factual statements that are not based on evidence or source. Evidence/arguments are not only false but could be completely fabricated to appear "real".
 4. **ASSOCIATION** - Making your audience experience a simplistic, one-sided emotional response to a complex event through words, images, or testimony related person that evoke strong positive/negative feelings for the target to promote the interests of one of the parties.
 5. **MALIGN RHETORIC** - Linguistic tricks aimed at silencing opinions. By undermining legitimate debate or by spreading malicious satire/gossip to circumvent the fact-checking institutions.
 6. **BANDWAGON EFFECT** - The information is presented to the target audience in order to convince them to join or take action that "everybody is going to take".

Chapter 3. Contributions

7. **CLICKBAIT** - A headline that uses sensational incentives that it does not reflect in the content.
- **Ad. 2.** EUvsDisinfo classification can be found in [120] and consists of:
 1. **THE STRAW MAN** - Attack views or ideas that the cited party has never expressed.
 2. **DENIAL** - Deny all allegations.
 3. **MOCKERY** - Using sarcasm to belittle the opposing party.
 4. **PROVOCATION** - A preventive rhetorical tool designed to frame a discussion and provoke a reaction.
 5. **ATTACK** - Using violent language to provoke an equally harsh reaction or to silence an opponent.
 6. **WHATABOUTISM** - Taking the discussion off-topic.
 7. **EXHAUSTION** - “Drowning” on the opposite side in technicalities and details.
 - **Ad. 3.** Medical disinformation categories:
 1. **SLIPPERY SLOPE** - The sentence is factually true, but the consequences of the presented fact are exaggerated.
 2. **ALLEGED NEGATIVE CONSEQUENCES** - The sentence is mostly true, but given the context of the expert’s experience, there is a risk that the information presented may induce the patient to act inconsistently with current medical guidelines.
 3. **TWISTING WORD** - The sentence seems to be true, but there is one word that changes its overtone.
 4. **CONSPIRACY THEORY** - The sentence has the hallmarks of a conspiracy theory.
 5. **HEDGING** - The sentence is factually untrue, but there is a part of it that softens the tone of the presented statement.
 6. **ANECDOTAL EVIDENCE** - A sentence contains a true statement based on a single example, often from personal or anecdotal experience.
 7. **MISLEADING STATISTICAL EVIDENCE** - Actually valid statistical evidence that is taken out of context or has misinterpretation.

Chapter 3. Contributions

DebunkEU	EUvsDisinfo	Medical	Aggregated
Hyperbolization	-	Slippery Slope Alleged negative consequences Misleading statistical evidence	HYPERBOLIZATION
Selection	-	Misleading statistical evidence	MISLEADING CONTEXT
Forgery	The straw man Denial	Twisting word	FACT MANIPULATION
-	-	Conspiracy theory	APPEALING TO INTEREST GROUPS
Association		Anecdotal evidence	SIMPLIFICATION
Malign rethoric	Mockery Provocation Attack	-	MALIGN ACTIONS
Bandwagon effect	-	Conspiracy theory	CALL TO ACTION
ClickBait	-	Slippery slope Conspiracy theory	SENSATION MANAGEMENT
-	Whataboutism Exhaustion	Hedging	OVERTALKING THE TRUTH

Table 3.1: Disinformation categories proposed by different media and the aggregated category labels

According to my observations, disinformation on medical topics has the following unique characteristics:

- conspiracy theories are more common in the medical domain, and the related ideas are less aggressively presented than in the general news domain. Related rhetoric is based on manipulations other than a strong appeal to emotions.
- some categories are not found anywhere else, such as unintentionally false, outdated information.
- To differentiate between some categories expert’s experience is needed, e.g., ‘alleged negative consequences.’

3.4 Contribution 4. Topical classifiers of credibility of medical sentence triplets with 90% precision in credible class

This contribution relates to Objective 6: **Designing filtering methods for the chosen textual units.**

Per-topic classifiers were constructed as credibility filters for sentence triplets. The features taken into consideration included:

1. Uncased TF-IDF (“word-count” or non-compressed lexical features) or BioBERT vectors (compressed lexical features)
2. Dependency tree-labels count (stylometric features)
3. Named entities count (stylometric features)
4. Polarity and subjectivity (stylometric features)
5. LIWC (stylometric features)

Before training each model, the Recursive Feature Elimination feature selection was performed. Then, a genetic algorithm was used to choose the best model and its hyperparameters. As it is stated in the description of the fourth research

objective, the credibility classification task should be assessed based on two criteria, which were both met:

1. **Precision for the positive (credible) class** - the classifiers achieve high Precision exceeding 90% for most medical topics considered in our study (vaccination, allergy testing, children antibiotics, steroids for kids, antioxidants, cholesterol & statins, and C-section vs. natural birth)
2. **Proportion of non-credible messages in the negative (non-credible) class (Negative predictive value), as opposed to their proportion in the non-filtered dataset** - for all the topics, the improvement in the utilization of medical experts' time is substantially better, with an average improvement of 25.9 percentage points, which means that within the same amount of time and at the same average time needed to annotate a single sentence, medical experts using our method annotate over twice as many non-credible medical statements, on average

A detailed description of the data augmentation techniques, features, feature selection methods, and models fine-tuning are elaborated on in the article **Focus On Misinformation: Improving Medical Experts' Efficiency Of Misinformation Detection** which was presented during the **International Conference on Web Information Systems Engineering 2021** conference and published in "Lecture Notes in Computer Science" (LNCS, volume 13081).

3.5 Contribution 5. Comparison of two classification methods in terms of their generalization and explanatory power

This contribution is related to Objective 7: **Interpreting and explaining decisions made by classification models.**

Classifiers from the previous study [115] were used to interpret the filtering criteria for medical message credibility. I found that two classes of classifiers

Chapter 3. Contributions

perform equally well, but provide distinct kinds of feedback:

1. **Classifiers based on non-compressed semantic features and stylometric features** - the proportion of stylometric features from the sets of the most important model features for each sub-domain is low and in favor of semantic features. Depending on the topic, semantic features hold from 85.3% to 96.1% share. Although the appearance of specific terms regarding e.g. unproven therapies given the narrow topic may be easy to interpret by a lay user, such a large share of semantic features diminishes the model's generalization ability as stylometric features are topic independent.
2. **Classifiers based on neural word embeddings and stylometric features** - the compression of the single words' or terms' meaning using a state-of-the-art language model over the set of sentences (in my example BioBERT performed best) significantly reduces the share of semantic features for the credibility classifiers. Semantic features hold from 82.9% to as small as 50.0% share, depending on the topic. These classifiers provide a better generalization but make semantic features impossible to interpret.

As a result of the study that aimed to reach Research Objective 7, it was shown that there is a trade-off between the medical credibility classifier's generalization ability and its explainability. For a detailed description of the explainable models (Logistic Regression with feature weights) and neural models with local explanations, please refer to the article **Improving medical experts' efficiency of misinformation detection: an exploratory study** published in **World Wide Web (2022) (IF 3.0)** as part of the collection **Special Issue on Web Information Systems Engineering 2021**.

Discussion

4.1 Concluding remarks

Using the credibility classifiers can be regarded as an initial filter for medical Web content. In a realistic use-case scenario, medical experts would continually evaluate a stream of statements derived from the ever-growing set of online articles on medical and health topics and information from social media. Filtering classifiers will increase the efficiency of misinformation detection by medical experts, who will discover more than twice as much misinformation without increasing either the time spent on the evaluation or the number of evaluating experts and without any changes to the annotation workflow. Moreover, I showed that modifying the input features could provide end-users with different types of feedback, either semantic or stylometric, without any performance loss. Because classifiers cannot provide semantic and stylometric explanations, it remains to be examined which type of feedback is more beneficial.

The outcome of my work was designed to facilitate the creation of a crowd-sourced, expert-in-the-loop credibility assessment system. The system could be supported by algorithms proposed in my research in the following ways:

- credibility classifiers for sentence triplets can pre-filter the data to remove triplets evaluated as credible with high certainty. It leads to an increase in the system's throughput (the amount of information evaluated by experts or the crowd in a unit of time)
- explainable credibility classifiers can help experts investigate the reasons for

the system's filtering decisions, reducing the error rate

Another valuable benefit of my research is detecting narrative or manipulation techniques in disinformation on medical topics. The qualitative analysis described in [114] has revealed four distinct narratives in non-credible sentences. Although the analysis was limited to cholesterol and statins, it is safe to suggest that these narratives are more general and may broadly apply to false medical information on other topics.

4.2 Limitations

One of the disadvantages of the proposed solution is using sentence triplets as a unit for credibility assessments. The medical community has become accustomed to content quality assessment protocols that consider the entire online articles. These tools are mature and tested. For the proposed unit, it is necessary to create and adapt an annotation protocol and start implementing classification models from scratch. However, as already mentioned, assessments of sentence triplets are easier to obtain in a distributed way. Moreover, classification models built upon sentence triplets can provide more precise explanations of their decisions, taking into account, for example, the rhetorical schemes used.

Another limitation of the proposed methods of data pre-filtering is that, due to false positive samples, a certain number of statements that contain misinformation would not be identified by experts. However, we need to remember that medical experts may not spot all the statements anyway, as their limited time and attention prevent them from processing all the suspicious information.

The mere assumption of obtaining data from medical experts using crowdsourcing may be too much of a challenge. Dependence on the participation of human experts is a general drawback attributed to Interactive Machine Learning models. The revolution in artificial intelligence, as stated in [1] "was largely based on taking humans out of the equation in exchange for substantially increasing computational requirements," but "IML systems promise to lower these computational requirements and make learning more efficient, in exchange for bringing humans

back into the equation with the problems associated to them (availability, attention, interactivity, different expertise, etc.)". Attracting experts to "do the job" is one of the aspects that is worth devoting to a separate branch of research. Gamification methods, specialized training for medical students, financial gratification, and a large dispersion of experts are some ideas for real-world system implementation worth testing and verifying.

4.3 Future work

My future efforts will be focused on fulfilling research Objective 4: improving the annotation protocol to obtain a high inter-rater agreement. Reaching this goal requires several iterations of improvements, following rounds of annotation events, calculating agreements, and qualitative data examination. After obtaining a better protocol and more data, filtering classifiers can improve further.

Another path of my future research activity is to focus on gathering more data by introducing the demo expert crowd-sourcing system at selected medical universities. I also plan to prepare professional training for medical students. The goal is to elevate medical students' annotation accuracy to the expert level (like medical practitioners with at least a few years of experience), thus reducing the costs of expert medical credibility annotation.

To develop and further extend the explainable power of filtering classifiers, designing machine learning models for detecting narrative forms [114] may be an exciting research direction (e.g., a model searching for instances of hedging expressions or words capable of twisting the overtone of the sentence). Tagging these narratives during credibility annotation may increase the precision of sentence classifiers built upon such datasets and, most importantly, help to disambiguate the experts' labeling process and improve the protocol.

My work on the annotation protocols has revealed that experts, similarly to lay users, are susceptible to various errors in their credibility evaluation. Research in psychology has revealed the role of cognitive biases in such errors. A better understanding of the impact of cognitive biases on credibility evaluations is needed.

Chapter 4. Discussion

It can improve the credibility evaluation processes, both for experts and for lay users. In particular, interesting questions for future research are: which categories of disinformation are the most susceptible to cognitive biases? Do different categories of disinformation trigger different cognitive biases?

In order to formulate this question in the form of a hypothesis, I conducted a review of research in psychology on cognitive biases that can impact credibility evaluation. After creating a universal categorization of disinformation in Section 3.3 I selected cognitive biases and other phenomena that I consider related to particular categories of disinformation. For each category, I attempted to select a cognitive bias that, in my view, is mainly responsible for the effectiveness of this disinformation category.

The selected biases and phenomena per category are presented in Table 4.1 and explained below:

- **Information overload** - is the difficulty in understanding an issue and effectively making decisions when one has too much information about that issue [128].
- **Omission bias** - Omission bias is "people's tendency to evaluate harm done through omission as less morally wrong and less blameworthy than commission when there is harm." [121]. In the context of medical disinformation, it corresponds to the aggregated category HYPERBOLIZATION. A great deal of medical disinformation that belongs to this category indicates the adverse effects of certain therapies making them appear more severe than in reality (with a vivid example being vaccinations). People's general inability to intuitively calculate probabilities has been extensively studied for decades, considering Tversky and Kahneman's famous publications alone [129]. This, together with the Omission bias, makes hyperbolization of a therapy's adverse effects an extremely dangerous tool for manipulation.
- **Contrast effect** - according to [123] is "the enhancement or diminishment, relative to normal, of perception, cognition, or related performance as a result of successive (immediately previous) or simultaneous exposure to a stimulus of lesser or greater value in the same dimension." A factually correct message when intentionally set against a very positive/negative context (that may

Chapter 4. Discussion

Aggregated category	Phenomena in the medical domain	Phenomena in the general news domain
HYPERBOLIZATION	Omission bias [121]	Emotional see-saw [122]
MISLEADING CONTEXT	Contrast effect (a subgroup of the context effects) [123]	Contrast effect
FACT MANIPULATION	Mere-exposure effect [124]	Mere-exposure effect
APPEALING TO INTEREST GROUPS	Out-group homogeneity [125], Confirmation bias	Out-group homogeneity, Emotional see-saw
SIMPLIFICATION	Cognitive dissonance reduction [126]	Cognitive dissonance reduction
MALIGN ACTIONS	Emotional see-saw	Emotional see-saw
CALL TO ACTION	The principle of commitment and consistency [127]	The principle of commitment and consistency
SENSATION MANAGEMENT	Emotional see-saw	Emotional see-saw
OVERTALKING THE TRUTH	Information overload	Information overload

Table 4.1: Aggregated disinformation categories and psychological biases and effects that correspond to them

even be fabricated) will drastically change its overtone.

- **Mere-exposure effect** - is a psychological phenomenon by which people tend to develop a preference for things merely because they are familiar with them [124]. An example of the effect appears intuitive: people will likely feel less anxious about the given piece of information that contains some fabricated, untrue statements when being exposed to them often. Like a Goebbels lie which repeated a hundred times will ultimately become perceived as truth.
- **Out-group homogeneity effect** is the perception of out-group members as more similar to one another than in-group members, e.g. "they are alike; we are diverse" [125]. This statement is the foundation for conspiracy theories and a leading tool to polarize societies.
- **Cognitive dissonance** - In the study from 1957 [130], Leon Festinger proposed that human beings strive for internal psychological consistency to function well mentally in the real world. The more complex a given phenomenon is, the harder it is for the reader to follow the story that describes it without a feeling of psychological discomfort. To ease this burden, misinformation creators propose a soft, simplified story or an individual's perspective. All of them finally miss the important context and are likely to change the overtone of the message.
- **Emotional see-saw** - is a technique of social influence. This mechanism consists in introducing a person into a state of fear, then, after a sudden and unexpected withdrawal of the negative stimulus, relief occurs, which is accompanied by a state of unreflectiveness [122]. This phenomenon may be especially dangerous when applied as part of highly sensational news with the following call for action.
- **The principle of commitment and consistency** - in disinformation campaigns this effect may be used as a way to make the audience less likely to change their views, even after being exposed to convincing yet contradictory arguments. People's tendency to behave in a manner that matches their past decisions or behaviors [127] after any involvement in an action to which they had been called, causes too great a cognitive effort to change their views.
- **Confirmation bias** is a tendency to process information by looking for, or interpreting, information that is consistent with one's current beliefs [131]

Chapter 4. Discussion

I attempted to synthesize the current research results in two areas: categorization and description of disinformation and cognitive biases. I propose a relationship exists between susceptibility to particular disinformation categories and cognitive biases or other psychological phenomena. The relationship proposed in Table 4.1 summarizes my hypotheses. Rigorous verification of these hypotheses can lead to an increased understanding of credibility evaluation and an improvement of systems for credibility evaluation support through a design that will be more resistant to cognitive biases.

Bibliography

- [1] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Fernández-Leal, “Human-in-the-loop machine learning: a state of the art,” *Artificial Intelligence Review*, 8 2022.
- [2] R. Fiebrink, P. R. Cook, and D. Trueman, “Human model evaluation in interactive supervised learning,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (New York, NY, USA), pp. 147–156, ACM, 5 2011.
- [3] A. Wierzbicki, *Web Content Credibility*. Cham: Springer International Publishing, 2018.
- [4] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, “The science of fake news,” *Science*, vol. 359, pp. 1094–1096, 3 2018.
- [5] A. L. Ginsca, A. Popescu, and M. Lupu, “Credibility in Information Retrieval,” *Foundations and Trends® in Information Retrieval*, vol. 9, no. 5, pp. 355–475, 2015.
- [6] C. I. Hovland, I. L. Janis, and H. H. Kelley, *Communication and persuasion; psychological studies of opinion change*. New Haven, CT, US: Yale University Press, 1953.
- [7] C. L. Corritore, S. Wiedenbeck, B. Kracher, and R. P. Marble, “Online Trust and Health Information Websites,” *International Journal of Technology and Human Interaction*, vol. 8, pp. 92–115, 10 2012.
- [8] C. L. Corritore, S. Wiedenbeck, and B. Kracher, “The elements of online trust,” in *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, (New York, NY, USA), pp. 504–505, ACM, 3 2001.

Bibliography

- [9] D. Artz and Y. Gil, “A survey of trust in computer science and the Semantic Web,” *Journal of Web Semantics*, vol. 5, pp. 58–71, 6 2007.
- [10] T. M. Ciolek, “The six quests for the electronic grail: Current approaches to information quality in WWW resources,” *Extrait de la Revue Informatique et Statistique dans les Sciences humaines*, XXXII, 1996.
- [11] C. Castelfranchi, R. Falcone, and G. Pezzulo, “Trust in information sources as a source for trust,” in *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, (New York, NY, USA), pp. 89–96, ACM, 7 2003.
- [12] M. Kąkol, M. Jankowski-Lorek, K. Abramczuk, A. Wierzbicki, and M. Catasta, “On the subjectivity and bias of web content credibility evaluations,” in *Proceedings of the 22nd International Conference on World Wide Web*, (New York, NY, USA), pp. 1131–1136, ACM, 5 2013.
- [13] R. Nielek, A. Wawer, M. Jankowski-Lorek, and A. Wierzbicki, “Temporal, Cultural and Thematic Aspects of Web Credibility,” in *SocInfo 2013: Social Informatics*, pp. 419–428, Springer, 2013.
- [14] S. Y. Syn and S. U. Kim, “The impact of source credibility on young adults’ Health information activities on facebook: Preliminary findings,” *Proceedings of the American Society for Information Science and Technology*, vol. 50, no. 1, pp. 1–4, 2013.
- [15] S. Stoerger, “I’m not a doctor, but I play one on the web: Credibility, funding, and interactivity features on health organization websites,” *Proceedings of the American Society for Information Science and Technology*, vol. 44, pp. 1–5, 10 2008.
- [16] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, “Propagation of trust and distrust,” in *Proceedings of the 13th international conference on World Wide Web*, (New York, NY, USA), pp. 403–412, ACM, 5 2004.
- [17] M. J. Metzger, A. J. Flanagin, and R. B. Medders, “Social and Heuristic Approaches to Credibility Evaluation Online,” *Journal of Communication*, vol. 60, pp. 413–439, 8 2010.
- [18] P. Borzysmek and M. Sydow, “Trust and Distrust Prediction in Social Network with Combined Graphical and Review-Based Attributes,” in *KES-*

Bibliography

- AMSTA 2010: Agent and Multi-Agent Systems: Technologies and Applications*, pp. 122–131, Springer, 2010.
- [19] A. Juffinger, M. Granitzer, and E. Lex, “Blog credibility ranking by exploiting verified content,” in *Proceedings of the 3rd workshop on Information credibility on the web*, (New York, NY, USA), pp. 51–58, ACM, 4 2009.
- [20] K. Lorek, J. Suehiro-Wiciński, M. Jankowski-Lorek, and A. Gupta, “Automated credibility assessment on Twitter,” *Computer Science*, vol. 16, no. 2, p. 157, 2015.
- [21] S. Shepperd and D. Charnock, “DISCERN,” *Health Expectations*, vol. 1, pp. 134–135, 11 1998.
- [22] Health On the Net, “About Health On the Net.” <https://www.hon.ch/en/about.html>. Accessed: 1 2023.
- [23] M. García Lozano, J. Brynielsson, U. Franke, M. Rosell, E. Tjörnhammar, S. Varga, and V. Vlassov, “Veracity assessment of online data,” *Decision Support Systems*, vol. 129, p. 113132, 2 2020.
- [24] S. Wiegand and S. E. Middleton, “Veracity and Velocity of Social Media Content during Breaking News,” in *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, (New York, New York, USA), pp. 751–756, ACM Press, 2016.
- [25] Y. Namihira, N. Segawa, Y. Ikegami, K. Kawai, T. Kawabe, and S. Tsuruta, “High Precision Credibility Analysis of Information on Twitter,” in *2013 International Conference on Signal-Image Technology & Internet-Based Systems*, pp. 909–915, IEEE, 12 2013.
- [26] S. Jain, V. Sharma, and R. Kaushal, “Towards automated real-time detection of misinformation on Twitter,” in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2015–2020, IEEE, 9 2016.
- [27] A. Vlachos and S. Riedel, “Fact Checking: Task definition and dataset construction,” in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, (Stroudsburg, PA, USA), pp. 18–22, Association for Computational Linguistics, 2014.
- [28] N. Hassan, C. Li, and M. Tremayne, “Detecting Check-worthy Factual Claims in Presidential Debates,” in *Proceedings of the 24th ACM International on*

Bibliography

- Conference on Information and Knowledge Management*, (New York, NY, USA), pp. 1835–1838, ACM, 10 2015.
- [29] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, “Computational Fact Checking from Knowledge Networks,” *PLOS ONE*, vol. 10, p. e0128193, 6 2015.
- [30] A. Magdy and N. Wanas, “Web-based statistical fact checking of textual documents,” in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, (New York, NY, USA), pp. 103–110, ACM, 10 2010.
- [31] B. Shi and T. Wenginger, “Fact Checking in Heterogeneous Information Networks,” in *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, (New York, New York, USA), pp. 101–102, ACM Press, 2016.
- [32] A. Tchechmedjiev, P. Fafalios, K. Boland, M. Gasquet, M. Zloch, B. Zapilko, S. Dietze, and K. Todorov, “ClaimsKG: A Knowledge Graph of Fact-Checked Claims,” in *ISWC 2019: The Semantic Web – ISWC 2019*, pp. 309–324, Springer, 2019.
- [33] N. Vedula and S. Parthasarathy, “FACE-KEG: Fact Checking Explained using KnowledgeE Graphs,” in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, (New York, NY, USA), pp. 526–534, ACM, 3 2021.
- [34] G. Levchuk and E. Blasch, “Probabilistic graphical models for multi-source fusion from text sources,” in *2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, pp. 1–10, IEEE, 5 2015.
- [35] P. Shiralkar, A. Flammini, F. Menczer, and G. L. Ciampaglia, “Finding Streams in Knowledge Graphs to Support Fact Checking,” in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 859–864, IEEE, 11 2017.
- [36] Z. Kou, L. Shang, Y. Zhang, C. Youn, and D. Wang, “FakeSens: A Social Sensing Approach to COVID-19 Misinformation Detection on Social Media,” in *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pp. 140–147, IEEE, 7 2021.

- [37] L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee, “DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (New York, NY, USA), pp. 492–502, ACM, 8 2020.
- [38] T. Karmakharm, N. Aletras, and K. Bontcheva, “Journalist-in-the-Loop: Continuous Learning as a Service for Rumour Analysis,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, (Stroudsburg, PA, USA), pp. 115–120, Association for Computational Linguistics, 2019.
- [39] T.-D. Cao, L. Duroyon, F. Goasdoué, I. Manolescu, and X. Tannier, “BeLink: Querying Networks of Facts, Statements and Beliefs,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, (New York, NY, USA), pp. 2941–2944, ACM, 11 2019.
- [40] M. Amith and C. Tao, “Representing vaccine misinformation using ontologies,” *Journal of Biomedical Semantics*, vol. 9, p. 22, 12 2018.
- [41] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, and J. G. Simonsen, “MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (Stroudsburg, PA, USA), pp. 4684–4696, Association for Computational Linguistics, 2019.
- [42] W. Y. Wang, ““Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Stroudsburg, PA, USA), pp. 422–426, Association for Computational Linguistics, 2017.
- [43] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, “Fighting an Infodemic: COVID-19 Fake News Dataset,” in *CONSTRAINT 2021: Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pp. 21–29, Springer, 11 2021.

- [44] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: a Large-scale Dataset for Fact Extraction and VERification,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (Stroudsburg, PA, USA), pp. 809–819, Association for Computational Linguistics, 2018.
- [45] G. K. Shahi, J. M. Struß, and T. Mandl, “Overview of the CLEF-2021 CheckThat! Lab: Task 3 on Fake News Detection,” in *CLEF (Working Notes)*, pp. 406–423, 2021.
- [46] M. Jankowski-Lorek, R. Nielek, A. Wierzbicki, and K. Zieliński, “Predicting Controversy of Wikipedia Articles Using the Article Feedback Tool,” in *Proceedings of the 2014 International Conference on Social Computing*, (New York, NY, USA), pp. 1–7, ACM, 8 2014.
- [47] B. Balcerzak and W. Jaworski, “Application of linguistic cues in the analysis of language of hate groups,” *Computer Science*, vol. 16, no. 2, p. 145, 2015.
- [48] S. Shaar, M. Hasanain, B. Hamdan, Z. S. Ali, F. Haouari, A. Nikolov, M. Kutlu, Y. S. Kartal, F. Alam, and G. Da San Martino, “Overview of the CLEF-2021 CheckThat! Lab Task 1 on Check-Worthiness Estimation in Tweets and Political Debates,” in *CLEF (Working Notes)*, pp. 369–392, 2021.
- [49] V. Wagle, K. Kaur, P. Kamat, S. Patil, and K. Kotecha, “Explainable AI for Multimodal Credibility Analysis: Case Study of Online Beauty Health (Mis)-Information,” *IEEE Access*, vol. 9, pp. 127985–128022, 2021.
- [50] N. Sitaula, C. K. Mohan, J. Grygiel, X. Zhou, and R. Zafarani, “Credibility-Based Fake News Detection,” in *Disinformation, Misinformation, and Fake News in Social Media*, pp. 163–182, Springer, 2020.
- [51] W. Jaworski, E. Rejmund, and A. Wierzbicki, “Credibility Microscope: Relating Web Page Credibility Evaluations to Their Textual Content,” in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp. 297–302, IEEE, 8 2014.
- [52] R. Manjula and M. S. Vijaya, “Deep Neural Network for Evaluating Web Content Credibility Using Keras Sequential Model,” in *Advances in Electrical and Computer Technologies*, pp. 11–19, Springer, 2020.

Bibliography

- [53] N. Kovachevich, “Nkovachevich at CheckThat! 2021: BERT fine-tuning approach to fake news detection,” in *CLEF (Working Notes)*, pp. 537–544, 2021.
- [54] C.-G. Cusmuluc, M. A. Amarandei, I. Pelin, V.-I. Cociorva, and A. Iftene, “UAICS at CheckThat! 2021: Fake news detection,” in *CLEF (Working Notes)*, pp. 494–507, 2021.
- [55] F. Balouchzahi, H. L. Shashirekha, and G. Sidorov, “MUCIC at Check-That! 2021: FaDo-Fake News Detection and Domain Identification using Transformers Ensembling,” in *CLEF (Working Notes)*, pp. 455–464, 2021.
- [56] J. Pennington, R. Socher, and C. Manning, “Glove: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Stroudsburg, PA, USA), pp. 1532–1543, Association for Computational Linguistics, 2014.
- [57] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North*, (Stroudsburg, PA, USA), pp. 4171–4186, Association for Computational Linguistics, 2019.
- [58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 7 2019.
- [59] B. Moulton, L. S. Franck, and H. Brady, “Ensuring Quality Information for Patients: development and preliminary validation of a new instrument to improve the quality of written health care information,” *Health Expectations*, vol. 7, pp. 165–175, 6 2004.
- [60] M. Bunge, I. Mühlhauser, and A. Steckelberg, “What constitutes evidence-based patient information? Overview of discussed criteria,” *Patient Education and Counseling*, vol. 78, pp. 316–328, 3 2010.
- [61] Working Group GPGI, “Good practice guidelines for health information,” *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, vol. 110-111, pp. e1–e8, 2016.

Bibliography

- [62] A. Keselman, C. Arnott Smith, A. C. Murcko, and D. R. Kaufman, “Evaluating the Quality of Health Information in a Changing Digital Ecosystem,” *Journal of Medical Internet Research*, vol. 21, p. e11129, 2 2019.
- [63] D. Charnock, S. Shepperd, G. Needham, and R. Gann, “DISCERN: an instrument for judging the quality of written consumer health information on treatment choices,” *Journal of Epidemiology & Community Health*, vol. 53, pp. 105–111, 2 1999.
- [64] C. L. A. Clarke, S. Rizvi, M. D. Smucker, M. Maistro, and G. Zuccon, “Overview of the TREC 2020 Health Misinformation Track,” in *TREC*, 2020.
- [65] TREC, “Health Misinformation Track Assessing Guidelines.” <https://trec-health-misinfo.github.io/docs/AssessingGuidelines-2020.pdf>. Accessed: 1 2023.
- [66] L. Kinkead, A. Allam, and M. Krauthammer, “AutoDiscern: rating the quality of online health information with hierarchical encoder attention-based neural networks,” *BMC Medical Informatics and Decision Making*, vol. 20, p. 104, 12 2020.
- [67] Z. Shah, D. Surian, A. Dyda, E. Coiera, K. D. Mandl, and A. G. Dunn, “Automatically Appraising the Credibility of Vaccine-Related Web Pages Shared on Social Media: A Twitter Surveillance Study,” *Journal of Medical Internet Research*, vol. 21, p. e14007, 11 2019.
- [68] G. Schwitzer, “How Do US Journalists Cover Treatments, Tests, Products, and Procedures? An Evaluation of 500 Stories,” *PLoS Medicine*, vol. 5, p. e95, 5 2008.
- [69] F. Afsana, M. A. Kabir, N. Hassan, and M. Paul, “Automatically Assessing Quality of Online Health Articles,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 591–601, 2 2021.
- [70] M. Al-Jefri, R. Evans, J. Lee, and P. Ghezzi, “Automatic Identification of Information Quality Metrics in Health News Stories,” *Frontiers in Public Health*, vol. 8, 12 2020.
- [71] I. B. Schlicht, A. F. M. de Paula, and P. Rosso, “UPV at TREC Health Misinformation Track 2021 Ranking with SBERT and Quality Estimators,” in *TREC*, 12 2021.

Bibliography

- [72] L. Cui and D. Lee, “CoAID: COVID-19 Healthcare Misinformation Dataset,” *arXiv*, 5 2020.
- [73] N. L. Kolluri and D. Murthy, “CoVerifi: A COVID-19 news verification system,” *Online Social Networks and Media*, vol. 22, p. 100123, 3 2021.
- [74] Y. Aphinyanaphongs and C. Aliferis, “Text categorization models for identifying unproven cancer treatments on the web,” *Studies in health technology and informatics*, vol. 129, no. Pt 2, pp. 968–72, 2007.
- [75] D. Hawking, T. Tang, K. M. Griffiths, N. Craswell, and P. Bailey, “Towards higher quality health search results: Automated quality rating of depression websites,” in *Proceedings of Medinfo 2007 Workshop on “Models of trust for health websites*, 2007.
- [76] Y. Wang, M. McKee, A. Torbica, and D. Stuckler, “Systematic Literature Review on the Spread of Health-related Misinformation on Social Media,” *Social Science & Medicine*, vol. 240, p. 112552, 11 2019.
- [77] Y. Zhao, J. Da, and J. Yan, “Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches,” *Information Processing & Management*, vol. 58, p. 102390, 1 2021.
- [78] R. Sicilia, S. Lo Giudice, Y. Pei, M. Pechenizkiy, and P. Soda, “Twitter rumour detection in the health domain,” *Expert Systems with Applications*, vol. 110, pp. 33–40, 11 2018.
- [79] M. H. Purnomo, S. Sumpeno, E. I. Setiawan, and D. Purwitasari, “Keynote Speaker II: Biomedical Engineering Research in the Social Network Analysis Era: Stance Classification for Analysis of Hoax Medical News in Social Media,” *Procedia Computer Science*, vol. 116, pp. 3–9, 2017.
- [80] A. Ghenai and Y. Mejova, “Fake Cures: User-centric Modeling of Health Misinformation in Social Media,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, pp. 1–20, 11 2018.
- [81] Z. Xu and H. Guo, “Using Text Mining to Compare Online Pro- and Anti-Vaccine Headlines: Word Usage, Sentiments, and Online Popularity,” *Communication Studies*, vol. 69, pp. 103–122, 1 2018.

Bibliography

- [82] H. Samuel and O. Zaïane, “MedFact: Towards Improving Veracity of Medical Information in Social Media Using Applied Machine Learning,” in *Canadian AI 2018: Advances in Artificial Intelligence*, pp. 108–120, Springer, 2018.
- [83] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil, “People on Drugs: Credibility of User Statements in Health Communities,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 65–74, ACM, 8 2014.
- [84] F. M. Dito, H. A. Alqadhi, and A. Alasaadi, “Detecting Medical Rumors on Twitter Using Machine Learning,” in *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*, pp. 1–7, IEEE, 12 2020.
- [85] Y. Liu, K. Yu, X. Wu, L. Qing, and Y. Peng, “Analysis and Detection of Health-Related Misinformation on Chinese Social Media,” *IEEE Access*, vol. 7, pp. 154480–154489, 2019.
- [86] S. Dhoju, M. Main Uddin Rony, M. Ashad Kabir, and N. Hassan, “Differences in Health News from Reliable and Unreliable Media,” in *Companion Proceedings of The 2019 World Wide Web Conference*, (New York, NY, USA), pp. 981–987, ACM, 5 2019.
- [87] Q. Liu, Z. Zhang, Y. Li, T. Liu, D. Li, and J. Shi, “ICNet: Incorporating Indicator Words and Contexts to Identify Functional Description Information,” in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 7 2019.
- [88] X. Liu, B. Zhang, A. Susarla, and R. Padman, “YouTube for Patient Education: A Deep Learning Approach for Understanding Medical Knowledge from User-Generated Videos,” *ArXiv Computer Science*, 7 2018.
- [89] Z. Wang, Z. Yin, and Y. A. Argyris, “Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning,” *arXiv*, 12 2020.
- [90] J. Li, “Detecting False Information in Medical and Healthcare Domains: A Text Mining Approach,” in *Smart Health*, pp. 236–246, Springer, 2019.
- [91] E. Dai, Y. Sun, and S. Wang, “Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository,” *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 853–862, 1 2020.

Bibliography

- [92] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 1 2013.
- [93] O. Levy and Y. Goldberg, “Dependency-Based Word Embeddings,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Stroudsburg, PA, USA), pp. 302–308, Association for Computational Linguistics, 2014.
- [94] M. Nickel and D. Kiela, “Poincaré Embeddings for Learning Hierarchical Representations,” in *Advances in Neural Information Processing Systems 30*, 5 2017.
- [95] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 12 2017.
- [96] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 11 1997.
- [97] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Stroudsburg, PA, USA), pp. 1724–1734, Association for Computational Linguistics, 2014.
- [98] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in *31st Conference on Neural Information Processing Systems*, 6 2017.
- [99] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil, “Universal Sentence Encoder,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 3 2018.
- [100] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu,

Bibliography

- C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems 33*, 5 2020.
- [101] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 8 2019.
- [102] D. Chicco, “Siamese Neural Networks: An Overview,” *Methods in molecular biology (Clifton, N.J.)*, vol. 2190, pp. 73–94, 2021.
- [103] L. Logeswaran and H. Lee, “An efficient framework for learning sentence representations,” in *6th International Conference on Learning Representations*, 3 2018.
- [104] L. Wang, Z. Tu, A. Way, and Q. Liu, “Exploiting Cross-Sentence Context for Neural Machine Translation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (Stroudsburg, PA, USA), pp. 2826–2831, Association for Computational Linguistics, 2017.
- [105] P. Ren, Z. Chen, Z. Ren, F. Wei, J. Ma, and M. de Rijke, “Leveraging Contextual Sentence Relations for Extractive Summarization Using a Neural Attention Model,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (New York, NY, USA), pp. 95–104, ACM, 8 2017.
- [106] R. Kadlec, M. Schmid, O. Bajgar, and J. Kleindienst, “Text Understanding with the Attention Sum Reader Network,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 3 2016.
- [107] O. Vinyals, M. Fortunato, and N. Jaitly, “Pointer Networks,” in *Advances in Neural Information Processing Systems 28*, 6 2015.
- [108] F. Hill, A. Bordes, S. Chopra, and J. Weston, “The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations,” in *4th International Conference on Learning Representations*, 11 2015.
- [109] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman, “ConText: an algorithm for determining negation, experiencer, and temporal status

Bibliography

- from clinical reports,” *Journal of biomedical informatics*, vol. 42, pp. 839–51, 10 2009.
- [110] Y. Yamamoto, “Disputed Sentence Suggestion towards Credibility-Oriented Web Search,” in *APWeb 2012: Web Technologies and Applications*, pp. 34–45, Springer, 2012.
- [111] H. Shibuki, T. Nagai, M. Nakano, M. Ishioroshi, T. Matsumoto, and T. Mori, “Interactive Method for Generation of Mediatory Summary to Verify Credibility of Web Information,” *Journal of Natural Language Processing*, vol. 20, no. 2, pp. 75–103, 2013.
- [112] B. D. Horne, M. Gruppi, and S. Adali, “Trustworthy Misinformation Mitigation with Soft Information Nudging,” in *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pp. 245–254, IEEE, 12 2019.
- [113] A. Nabożny, B. Balcerzak, and D. Koržinek, “Enriching the Context: Methods of Improving the Non-contextual Assessment of Sentence Credibility,” in *Web Information Systems Engineering – WISE 2019*, pp. 763–778, Springer, 2019.
- [114] A. Nabożny, B. Balcerzak, A. Wierzbicki, M. Morzy, and M. Chlabicz, “Active Annotation in Evaluating the Credibility of Web-Based Medical Information: Guidelines for Creating Training Data Sets for Machine Learning,” *JMIR Medical Informatics*, vol. 9, p. e26065, 11 2021.
- [115] A. Nabożny, B. Balcerzak, M. Morzy, and A. Wierzbicki, “Focus on Misinformation: Improving Medical Experts’ Efficiency of Misinformation Detection,” in *Web Information Systems Engineering – WISE 2021*, pp. 420–434, Springer, 2021.
- [116] A. Nabożny, B. Balcerzak, M. Morzy, A. Wierzbicki, P. Savov, and K. Warpechowski, “Improving medical experts’ efficiency of misinformation detection: an exploratory study,” *World Wide Web*, 8 2022.
- [117] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Text,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, (Barcelona, Spain), pp. 404–411, Association for Computational Linguistics, 7 2004.

Bibliography

- [118] M. Ogrodniczuk and M. Lenart, “Web Service integration platform for Polish linguistic resources,” in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, (Istanbul, Turkey), pp. 1164–1168, European Language Resources Association (ELRA), 5 2012.
- [119] DebunkEU, “DebunkEU disinformation analysis, workshop materials,” 2020.
- [120] EUvsDisinfo.eu, “EUvsDisinfo methodology.” <https://euvsdisinfo.eu/modus-trollerandi-part-1-the-straw-man/>. Accessed: 1 2023.
- [121] S. K. Yeung, T. Yay, and G. Feldman, “Action and Inaction in Moral Judgments and Decisions: Meta-Analysis of Omission Bias Omission-Commission Asymmetries,” *Personality and Social Psychology Bulletin*, vol. 48, pp. 1499–1515, 10 2022.
- [122] D. Doliński and R. Nawrat, “Huśtawka emocji jako nowa technika manipulacji społecznej,” *Przegląd psychologiczny*, vol. 37, no. 1-2, pp. 16–19, 1994.
- [123] S. Plous, *The psychology of judgment and decision making*. McGraw-Hill series in social psychology., New York, NY, England: Mcgraw-Hill Book Company, 1993.
- [124] R. B. Zajonc, “Attitudinal effects of mere exposure,” *Journal of Personality and Social Psychology*, vol. 9, no. 2, Pt.2, pp. 1–27, 1968.
- [125] G. A. Quattrone and E. E. Jones, “The perception of variability within in-groups and out-groups: Implications for the law of small numbers,” *Journal of Personality and Social Psychology*, vol. 38, pp. 141–152, 1 1980.
- [126] L. Festinger, “Cognitive Dissonance,” *Scientific American*, vol. 207, pp. 93–106, 10 1962.
- [127] R. B. Cialdini, *Influence: Science and practice*. New York, NY, US: Harper-Collins College Publishers, 1993.
- [128] P. G. Roetzel, “Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development,” *Business Research*, vol. 12, pp. 479–522, 12 2019.
- [129] A. Tversky and D. Kahneman, “Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment,” *Psychological Review*, vol. 90, pp. 293–315, 10 1983.

Bibliography


- [130] L. Festinger, H. W. Riecken, and S. Schachter, *When prophecy fails: A social and psychological study of a modern group that predicted the destruction of the world*. New York: Harper & Row, 1964.
- [131] B. J. Casad and J. Luebering, “Confirmation bias.” <https://www.britannica.com/science/confirmation-bias>. Accessed: 1 2023.

CHAPTER 5

Articles comprising the thesis

- 5.1 Article 1 "Enriching the Context: Methods of Improving the Non-contextual Assessment of Sentence Credibility" (WISE 2019, 140 pts)

Enriching the Context: Methods of Improving the Non-contextual Assessment of Sentence Credibility

Aleksandra Nabożny¹ (✉) , Bartłomiej Balcerzak², and Danijel Koržinek²

¹ Gdańsk University of Technology, 80233 Gdańsk, Poland
alenaboz@pg.edu.pl

² Polish-Japanese Academy of Information Technology, 02008 Warsaw, Poland

Abstract. This paper presents several methods of automatic context enrichment of sentences that need to be evaluated, tagged or fact-checked by human judges. We have created a corpus of medical Web articles. Sentences from this corpus have been fact-checked by medical experts in two modes: contextually (reading the entire article and evaluating sentence by sentence) and without context (evaluating sentences from all articles in random order). It is known that non-contextual evaluation is faster, but some sentences are impossible to evaluate without context. We have designed and evaluated several methods of summarizing context that we hypothesized were suitable for supporting evaluation of sentences without reading the entire text. Then, we collected new assessments from medical experts for the sentences with enriched context. The context enrichment methods have been evaluated using two measures: conversion, which calculates how frequently a method allows experts to evaluate sentences that were impossible to evaluate without context, and agreement, which depends on how frequently the new expert evaluations match with evaluations from experts who had read the whole text before rating a sentence. Our results show that the best method achieves a high conversion rate, while providing experts with a condensed context summary. Moreover, the method significantly reduces the time needed to evaluate one sentence, compared to the baseline method (which provides the expert with the entire paragraph surrounding the target sentence). The problem of automatically enhancing the context of a sentence for fast fact-checking or tagging has not appeared in other studies before. We present preliminary results of the research in this area and a framework for testing potential new methods.

Keywords: Information credibility · Fact-checking · Text summarization · Context enrichment

1 Introduction

People often search the Web for medical or health-related information (“medical Web content” for short). As a matter of fact, eight in ten people browse the Web

for health-related content, which makes a consultation with “Dr Google” one of the most common Internet activities. Unfortunately, what the users ultimately find is often misleading, incomplete, and non-credible. The Web is filled with a myriad of humbug therapies, mysterious superdrugs and pseudo-doctors. As it can be easily guessed it may, and often does, lead to grave consequences, as can be seen by the example of the anti-vaccine movement, a global community that is largely present on the Web.

On the other hand, recent findings show that the trend of “googling” the diagnoses indeed permanently changes patient-doctor dialogue [2], with positive results to both sides. According to [20] it is likely that most people will experience at least one diagnostic error in their lifetime. Research suggests that the traditional diagnosis process is error-prone and it might be improved with some supporting methods, e.g. patient assistance in gathering knowledge about their own health condition.

Ordinary Web users’ credibility evaluation of medical Web content can be supported in several ways. Existing fact-checking sites, such as Snopes.com, have separate categories of non-credible medical Web content. Other sites, such as hon.ch, offer specialized search engines of medical Web content and run a certification process for medical Websites. Classifiers of Webpage credibility can be applied to augment output of Google search (using a browser plugin) with indications of search result’s credibility [26]. A more detailed method is to mark sentences contained in a medical Webpage with credibility indicators - research has shown that evaluations of statements can impact overall Webpage credibility evaluations [7,8]. This method also works for social media posts [19]. To obtain such information, evaluations of these sentences must be available. Thus, in this article we shall focus on supporting the process of acquiring sentence credibility evaluations from medical experts.

Consider a situation when a medical expert is asked to rate credibility of medical Web content. The most accurate, but also most time-consuming method would be for the expert to read the entire Webpage and to mark credibility of selected sentences. However, this method has a low output, because experts’ time is limited. Presenting experts with sentences selected for evaluation by Web users (or automatically) is another method. In this case, the medical expert only needs to read short sentences and can immediately give an evaluation. While this method has an advantage of speed, it has a drawback: contextual information that may be necessary for evaluation is missing.

It should be emphasized that in this study we limited ourselves to the medical field as a domain of research to reduce the amount and variety of necessary experts, keywords, etc. However, credibility assessment can be supported using the presented methods in any domain that requires an expert for the proper content evaluation. Other domains may include climate sciences, psychology, history, etc. Credibility evaluation is also important on Wikipedia, where teams of editors oversee quality and veracity of statements [22].

In order to cope with this challenge, we have designed and evaluated methods of context enrichment that help experts in assessing credibility of sentences

retrieved from medical Webpages. We also performed qualitative analysis of the sentences marked by experts as impossible to assess without context. We identified different types of such sentences and adapted the appropriate context-filling methods to each type. They are described in Sect. 4.

We asked medical experts to rate the credibility of sentences that have been found previously as impossible to evaluate without context, using each method separately. Results of the experimental evaluation are shown in Sect. 5. Our work bases on a dataset obtained in a previous experiment [14] (see Sect. 3), that is made available to interested researchers on request. The problem of automatically enriching the context in order to assess the credibility of a sentence has not appeared in any studies before. Note that this problem occurs whenever sentences need to be fact-checked quickly by experts, which means that it is significant for most kinds of fake news verification. Fast fact-checking is also crucial in the era of deep-learning models that need large sets of training data. Data that most oftenly needs to be tagged by human judges. We propose the first approach to the problem defined in this way, while opening the field for further research.

2 Related Work

Modelling of context is present in many solutions for downstream natural language processing tasks, however in most cases it serves only as an intermediate tool to enhance performance of these algorithms. These tasks include e.g. semantic text similarity, sentiment analysis and machine translation. The context is usually coded in a way that is not possible to interpret by a human, but only by a neural network that processes this context. In [12] surrounding sentences are used to better learn vector representations of the input sentence, similarly to the way that word2vec algorithm learns representation of the word. In [24], on the other hand, context summarized in a hierarchical way is integrated with neural machine translation model as a source for updating decoder states. In [18] the authors take advantage of contextual relations among sentences so as to improve the performance of sentence regression for text summarization.

In our study we aim to retrieve the context directly, which can be later accessed in a human readable format. A variety of methods exist that include direct extraction of context. Cloze-style reading comprehension problem has recently become a well known baseline NLP task. It is a task where the level of text understanding of a system is tested by asking it questions, the answer for which can be inferred from the document. In [6] the query is designed in a form of a short sentence that summarizes some statement which appears in a text, but lacks one named entity. Predicting the missing component requires a deep understanding of the context. The authors takes advantage of the popular deep-learning architectures with recurrent neural networks and attention to solve this problem. Their approach was to review and simplify the existing solutions, such as Pointer Networks [23] or Memory Neural Networks for text comprehension [5], which resulted as a new state of the art.

Aside from the aforementioned NLP methods, there is a whole other branch of methods which utilize rule-based algorithms for context extraction. These methods are used to support decision-making by retrieving context from electronic medical reports. For example ConText algorithm [4] derives information such as negation, experiencer and temporality of the medical condition. One of the methods presented in this study is also rule-based, but as an addition to the more general keyword-based approach.

Some of the previous works, designed to support credibility assessment of the query, take advantage of the automatically retrieved context. [27] uses global context (derived from the whole set of documents retrieved by the search engine) to prompt the user with sentences that may indicate controversy related to the given query, whereas [21] uses context to provide the information whether given article supports or rejects the statement contained in the query. Unlike in our approach, both studies are focused on the regular internet user (not an expert) and treat the query as a whole, not as part of a larger content.

3 Datasets

In this study we examine two datasets. One has been previously collected, analysed, and described in detail in [14] and we will focus only on one part of this dataset, which we will refer to as the first dataset. The second dataset has been collected especially for this study, using results from the first dataset.

The first dataset contains credibility assessments of articles retrieved from medical Webpages, as well as individual credibility evaluations of all sentences from those articles. The assessments were made by medical experts (doctors, Ph.D. students in medicine) on medical textual Web content (popular science articles addressed to a lay recipient). All articles were in Polish and have been assessed by medical experts who were Polish native speakers. This dataset was available for researchers upon request.

Experts evaluated individual sentences in one of the two possible procedures:

1. sentences in a given evaluation round were put one after another and formed the whole article (contextual mode),
2. sentences in a given round were taken at random from the whole corpus of articles (non-contextual mode).

Experts evaluated credibility of sentences marking them with one of the following labels:

- 0** - non-credible sentence,
- 1** - neutral sentence,
- 2** - credible sentence,
- 1** - impossible to assess due to the missing context (only in non-contextual mode).

This article focuses on the last case. In the reference study, during the first round, whole articles were additionally evaluated as either credible (2) or non-credible (0) regardless of the evaluation of individual sentences in the article. The full dataset summary is as follows:

- 247 evaluations of the whole articles (with only 0 or 2 label) collected,
- 11034 contextual evaluations of sentences collected, of which 3035 sentences were labelled as impossible to evaluate without context.

Interestingly, the percentage of sentences that were impossible to assess without context was 27.5% of the corpus of sentences. 24% of sentences marked as credible in the contextual mode were labelled as ‘-1’ in the non-contextual mode. Similarly, 24% of those marked as non-credible in the contextual mode were labelled as ‘-1’ in the non-contextual mode. Lastly, 36% of sentences marked as neutral in the contextual mode turned out to be impossible to assess without context in the non-contextual mode.

In this study we further investigate the last subset of sentences. We have performed an experiment in which 5 experts evaluated 500 randomly selected sentences from the subset of 3035 sentences that were impossible to assess in a non contextual mode. All those 500 sentences were grouped into the evaluation groups of 100 sentences. Each round consisted of 20 sentences with a pronoun, and 80 without pronouns, to reflect the distribution from the full set. 5 groups multiplied by 4 methods of context enrichment resulted in 2000 evaluations in total. The groups were sent to the respondents so that they would not encounter the same sentence twice (Fig. 1).

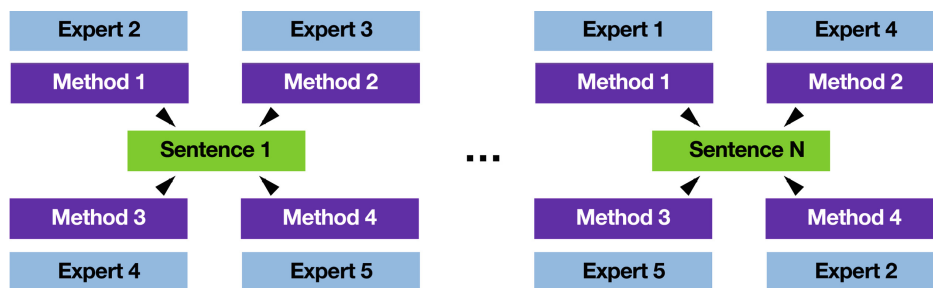


Fig. 1. Evaluation procedure: every sentence has its context enriched by all four methods designed throughout the study and is evaluated by different experts (every expert evaluates one sentence only once, but ultimately all experts use all methods on different sets of sentences).

Basing on the previous dataset we re-investigated the results for the purpose of this more focused study and we have found that experts had difficulty with assessing sentences that:

1. contained an “aggregate of meaning”, that is, a word being a hypernym that refers to a category of specific medical terms (for example, “This virus is dangerous” contains the word “virus”, which is considered a hypernym for the words “HIV”, “HPV” and “measles”),

2. contained an anaphoric, cataphoric or dialectical pronoun, eg. “They have serious consequences.”,
3. do not contain a subject (or subject is hard to identify),
4. that lacked a general context.

This analysis demonstrates the generality of the problem of missing context. This problem cannot be reduced to a single, well-known NLP problem.

4 Context Summarizing and Evaluation Methods

We have designed context summarizing methods described below, which we hypothesized were suitable for solving most problems with non-contextual sentence credibility evaluation. Experts were confronted with the sentences with added context summaries to compare (basing on the contextual evaluations from the first dataset) how the automatically extracted context changes their perception of the sentence.

4.1 Context Window (CW)

The first method consisted in retrieving the context window from the surrounding sentences. Two preceding sentences and one following sentence formed the context summary in this method. We treat this method as a baseline approach, in order to see to what extent the full, unprocessed information about the context is needed to correctly assess the credibility of a sentence.

Note that alternative methods were designed to produce shorter context summaries than the context window method. Sentence evaluation using the context window method would require significantly longer time from the evaluating experts, as shown in Sect. 5.

4.2 TF-IDF Keywords + Rule-Based Method of Supplementing the Meaning of Pronouns (TF-IDF + RB)

We used Term Frequency-Inverse Document Frequency (TF-IDF) statistic to retrieve 5 most relevant words from each article. We then attached them as keywords to the sentence that was to be evaluated. Then, for all sentences that contained pronouns, we applied the rule-based method of supplementing the meaning of pronouns, as described in Algorithm 1.

4.3 TextRank Keywords + Rule-Based Method of Supplementing the Meaning of Pronouns (TextRank + RB)

This method is a modification of the method described in Sect. 4.2. However, instead of calculating the Tf-Idf scores for words from the entire document, only 3 most relevant sentences were used. This method was used in order to focus on the most relevant parts of a document. This is important due to the fact

that some documents may be long. These sentences are selected based on the TextRank algorithm as described by [13], 3 sentences with the highest TextRank scores are selected. Next, the same rule-based method as described in Sect. 4.2 is used to complement the meaning of pronouns.

4.4 Rule-Based Method of Supplementing the Meaning of Pronouns

After we have identified sentences that consisted of anaphoric, cataphoric or dialectic pronouns, we used the Algorithm 1 to retrieve related noun phrase to the given pronoun. The longest considered noun phrase consists of a noun and a corresponding adjective. We used the Concraft [25] tool from Multiservice [16] web service for morphosyntactic tagging.

```
Result: noun [adjective]
INPUT: contextWindow (two preceding sentences) and a targetSentence (each sentence
is tokenized; each token has morphosyntactic tags attached);
for pronoun p in a targetSentence do
  candidate_target_nouns = all nouns in a target sentence that has a matching subset of
  {number, case, gender} with p ;
  if length of candidateTargetNouns  $\neq$  0 then
    | return first noun from the list ;
  else
    candidate_context_nouns = all nouns in a context sentences that has min. of 2
    overlapping values of {number, case, gender} with p ;
    if length of candidate_context_nouns  $\neq$  0 then
      | result = (if exist) first noun from the closer sentence (else) last noun from the
      | further sentence ;
      | if exists adjective in a 2-word context window for the resulting noun then
      |   | return result + adjective ;
      |   | else return result ;
    else
      | return empty string ;
    end
  end
end
```

Algorithm 1: Rule-based algorithm for supplementing the meaning of pronouns

4.5 Coreference Resolution (COREF)

The purpose of coreference resolution is to find words or groups of words that are linked to the same concept. These links can span over multiple sentences and serve as an important tool to explain the meaning, especially when analyzing fragments out of context.

In order to perform coreference resolution, we used the Multiservice [16] web service developed for the Polish language by researchers from the Clarin project (<https://clarin-pl.eu>). The coreference resolution pipeline consists of several tools activated in sequence:

1. Concraft [25] - initial segmentation and morphosyntactic tagging
2. Spejd [1] - morphosyntactic disambiguation and segment grouping
3. Nerf - named entity recognition

4. MentionStat [9] - detection of potential coreference candidate, so-called mentions
5. Bartek3 [10] - coreference resolution engine

The result of the whole process is a list of mention clusters. A mention is a segment or a sequence of segments representing the basic unit that can participate in a coreference. It can be a named entity or a term with an important semantic role, but it can also be a pronoun or another word that can easily be linked in the coreference. The mentions are grouped in clusters that form an equivalence relation between all the mentions in a given cluster. In theory, it would be possible to make the relations more meaningful by providing actual roles to the mentions, but the current version of the tool supports only equivalence relations.

The purpose behind using the tool was to enrich the context of sentences. In order to perform the coreference analysis, a set of short text fragments containing the mentioned sentences was prepared, with a context of two previous and one following sentence, giving four sentence per fragment. After collecting the list of mention clusters only those were kept that contained at least one mention within the analyzed sentence and at least one mention outside that sentence.

After collecting the coreferences for the whole data set, about 31% of the sentences didn't contain any mentions.

4.6 Performance Evaluation Measures

In this section, we present performance evaluation measures that will be used to evaluate the proposed methods of improving non-contextual sentence credibility assessment.

Conversion rate: This measure represents the percentage of sentences that had their credibility assessment changed from undetermined (in non-contextual evaluation) to either credible or not credible (after context enrichment). This measure shows the general efficiency of the method. The exact definition of the measure is provided by the following formula:

$$C = N_d / N_{n,d}$$

where N_d is the number of sentences which were given the evaluation (in place of the previous indeterminate assessment), while $N_{n,d}$ is the number of all sentences which had previously an indeterminate assessment.

Strong agreement rate: This measure represents the quality of the method for assessment improvement. It is a ratio of agreement between a non-contextual sentence evaluation (after the enhancement was applied), and the contextual assessment. A high ratio would indicate that the context transferring method was successful in recreating the context of the document in which the sentence

ORIGINAL SENTENCE:

(pl) Pojawiają się one dopiero w momencie, gdy organizm zakażonej osoby zaczyna walczyć wirusem i wytwarza przeciw niemu przeciwciała.

(eng) They appear only when the body of the infected person begins to fight the virus and makes antibodies against it.

TFIDF + RB

(pl) Słowa kluczowe artykułu: HIV zakazić wirus zakażenie test . Wybrane zdanie z artykułu: Pojawiają się one [Objawy] dopiero w momencie , gdy organizm zakażonej osoby zaczyna walczyć z wirusem i wytwarza przeciw niemu przeciwciała.

(eng) Article keywords: HIV infect virus infection test. Selected sentence from the article: They appear [Symptoms] only when the body of an infected person begins to fight the virus and makes antibodies against it.

TextRank + RB

(pl) Słowa kluczowe: zakazić HIV wirus test kobieta . Wybrane zdanie z artykułu: Pojawiają się one [Objawy] dopiero w momencie , gdy organizm zakażonej osoby zaczyna walczyć z wirusem i wytwarza przeciw niemu przeciwciała.

(eng) Keywords: infect HIV virus test woman. Selected sentence from the article: They appear [Symptoms] only when the body of the infected person begins to fight the virus and makes antibodies against it.

CW

(pl) Często może być tego faktu zupełnie nieświadoma, ponieważ infekcje mogą przebiegać przez długi czas bezobjawowo. Objawy HIV We wstępnej fazie zakażenia, kiedy HIV wnika do organizmu, żadne objawy nie są zauważalne ani odczuwalne. Pojawiają się one dopiero w momencie, gdy organizm zakażonej osoby zaczyna walczyć z wirusem i wytwarza przeciw niemu przeciwciała. Zakażony może wówczas czuć się tak, jak podczas grypy: będzie miał bęle głowy, mięśni i lekkie nabrzmienie węzów chłonnych.

(eng) She can often be completely unaware of this, because infections can be asymptomatic for a long time. In the initial stages of HIV infection, when HIV enters the body, no symptoms are noticeable or felt. They appear only when the body of an infected person begins to fight the virus and makes antibodies against it. The infected person may then feel as if they have the flu: she will have headaches, muscle aches and a slight swollen lymph nodes.

COREF

(pl) Pojawiają się one dopiero w [długi czas]momencie, gdy organizm zakażonej osoby zaczyna walczyć z wirusem i wytwarza przeciw niemu przeciwciała.

(eng) They appear only at a [long time] when the body of an infected person begins to fight the virus and makes antibodies against it.

Example 1.1. Exemplary sentence enriched with the presented context enrichment methods

was placed. Thus, it is a measure complementary to the conversion rate. The agreement rate is expressed by the following formula:

$$A = \frac{s_{agr}}{k} \quad (1)$$

where s_{agr} stands for the sum of all consistent pairs

$$s_{agr} = \sum_{n=1}^S 1 [s_{c,n} == s_{nce,n}] \quad (2)$$

s_c stands for contextual evaluation of the sentence, s_{nce} - enriched non-contextual evaluation, and k - number of pairs where s_{nce} does not equal -1

$$k = \sum_{n=1}^S 1 [s_{nce,n} \neq -1] \quad (3)$$

Weighted Agreement Rate: This measure is a modification of the strong agreement rate. We weigh the outcome of the assessments comparison so that:

1. the largest weight w_1 is assigned to the strong agreement (credible-credible and noncredible-noncredible assessment pairs between the contextual assessment mode and the non-contextual mode with context enrichment);
2. much smaller weight w_2 is assigned to all the pairs that contained “neutral” label on one side (either on C or NCE) (either on contextual or non-contextual with context enrichment). We justify this modification based on the assumption that misinterpretation of informative non-credible sentence with neutral, as well as informative credible with neutral, is potentially less harmful than misinterpretation of informative non-credible with informative credible. Moreover, there is much more randomness in assigning “neutral” labels to sentences by human judges than any other label type;
3. smaller weight w_3 , but closer to w_1 than to w_2 , is given to the neutral-neutral pairs, for the same reason related to the randomness of the assessments.
4. zero (weight w_4) is given in the strong disagreement scenario (credible-noncredible).

Ultimately, the weights are assigned as follows:

$$w = \begin{cases} 1, & \text{if } \{s_c, s_{nce}\} == \{2, 2\} \text{ or } \{0, 0\} \\ 0.33, & \text{if } \{s_c, s_{nce}\} == \{1, 2\} \text{ or } \{2, 1\} \text{ or } \{1, 0\} \text{ or } \{0, 1\} \\ 0.8, & \text{if } \{s_c, s_{nce}\} == \{1, 1\} \\ 0, & \text{otherwise} \end{cases}$$

And the formula is:

$$A_w = \frac{\sum_{n=1}^S w_n}{k} \quad (4)$$

where k , s_c , s_{nce} are defined as in Eqs. 2 and 3.

Standardized Length Factor. In the evaluation process we took into consideration the length factor of the retrieved context. It is expressed as a mean retrieved context length per method, divided by the mean length of the article. All lengths are represented as numbers of tokens.

$$LF = m_{cl}/m_{al} \quad (5)$$

where m_{cl} is the mean retrieved context length and m_{al} is the mean article length. We invert the LF because this is the factor that we want to minimize in the aggregate measure

$$ILF = 1 - LF \quad (6)$$

We standardize obtained values with min-max normalization function to the interval (0.5, 1) in order to make the outputs appear on the same scale as the other measures

$$SLF = MinMax(0.5, 1, min_{LF}, max_{LF}, v_{LF}) \quad (7)$$

MinMax is a linear transformation that takes as arguments: minimum and maximum value of the considered set, minimum and maximum value of the new interval, and the variable itself

$$\begin{aligned} & MinMax(min_{new}, max_{new}, min_{LF}, max_{LF}, v_{LF}) \\ &= \frac{v_{LF} - min_{LF}}{max_{LF} - min_{LF}} * (max_{new} - min_{new}) + min_{new} \end{aligned} \quad (8)$$

Weighted Harmonic Mean. In this paper, we claim that an optimal context enrichment method ought to maximize conversion rate and agreement rate, while at the same time minimize the length of the added context. Besides, the conversion rate is slightly less important than the agreement rate. That is why we introduced the summarizing measure: weighted harmonic mean. We consider this measure for both strong and weighted agreement rate. Eventually, the formula looks like follows:

$$WHM = \frac{3}{\frac{0.8}{C} + \frac{1.2}{A} + \frac{1}{SLF}} \quad (9)$$

5 Results and Discussion

Table 1 contains quality measures for all considered methods on the full dataset, while Table 2 limits the results to sentences that contained pronouns.

The Context Window method performs best when taking into consideration conversion, strong agreement and weighted agreement rates (as seen in Table 1). Surprisingly, the agreement rate of the Context Window method is only a few percent higher than the results for methods *TFIDF + RB* and *TextRank + RB*. Application of the length factor makes the Context Window method suboptimal, according to our criteria (maximizing C and A , and minimizing LF). TextRank

Chapter 5. Articles comprising the thesis

Table 1. Performance measures calculated for the full dataset of sentences with enriched context. Evaluations obtained with each method are compared to the fully contextual evaluations from the first dataset. C stands for Conversion rate, A - strong agreement rate, A_w - weighted agreement rate, SLF - standardized length factor, WHM_{SA} - weighted harmonic mean with strong agreement rate and WHM_{WA} - weighted harmonic mean with weighted agreement rate. All measures take values from 0 to 1, except SLF that takes values from 0.5 to 1.

	TFIDF + RB	TextRank + RB	CW	COREF
C	0.796	0.814	0.912	0.394
A	0.5	0.499	0.564	0.563
A_w	0.593	0.605	0.655	0.642
SLF	0.918	0.918	0.5	1
WHM_{SA}	0.67	0.677	0.599	0.581
WHM_{WA}	0.73	0.74	0.637	0.612

Keywords + rule-based algorithm for supplementing the meaning of pronouns appears to be optimal, taking into consideration both strong and weighted agreement rate. The results are close to those obtained by $TFIDF + RB$ method. It may indicate that shorter context, but collected from the full content of the article (as opposed to the context collected only from the surrounding sentences), proves to be sufficient for the expert to correctly assess the credibility of the sentence.

Table 2. Performance measures calculated only for the sentences that contained pronouns

	TFIDF + RB	TRK + RB	CW	COREF
C	0.84	0.83	0.93	0.54
A	0.44	0.506	0.548	0.352
A_w	0.569	0.591	0.665	0.491
SLF	0.918	0.918	0.5	1
WHM_{SA}	0.629	0.678	0.594	0.509
WHM_{WA}	0.723	0.735	0.643	0.609

We have also checked to what extent the methods for complementing the meaning of pronouns (both rule-based and the method resulting from finding coreferences) affect the overall score. We have selected and calculated measures only for the subset of sentences that contained a pronoun. While the conversion rate is significantly higher for all methods (especially for coreference resolution), it does not affect aggregate measures (as seen in Table 2).

We have performed chi-squared tests to check whether the evaluated context enrichment methods improve accuracy of credibility evaluation when compared

to random evaluations. In case of all methods, for both strong and weighted agreement rate, we can reject the null hypothesis that context enrichment does not improve accuracy at a 99% confidence level.

In general, the method of applying co-referent mentions to the sentences, at the current stage of model development, proved to be sub-optimal to the task of tagging the credibility of sentences. From studying the individual steps of the co-reference resolution pipeline, we did not notice any large-scale issues within the initial steps of the processing. We suspect that the main issue lies in the actual co-reference resolution engine Bartek. From what we can gather, the system is trained on the Polish Co-reference Corpus [15] which is hand-annotated co-reference corpus based largely on the Polish National Corpus [17] and other sources of news articles. The corpus does mention a very small percentage of scientific texts, but it is very unlikely it would contain any significant amount of medical texts.

In the course of the study we also experimented with the WordNet based method for completing the meaning of hyponymous expressions (eg. we tried to complement a word [virus] with [HIV], or [cancer] with [breast], [malignant]). We have collected the terminology from the article and compared it to the set of units that were linked in WordNet to the given word as hyperonyms. We faced the problem of too much generality in a WordNet structure and we decided to abandon this approach in favor of other methods. However, experiments with more domain-specific knowledge networks will be subject of our future work.

Reduction of Assessment Time. In the reference study [14] the Authors report the average article assessment time as approximately 10 min. Articles in the corpus have an average of 771 tokens, which should give about 3.5 min per article (given the average reading speed of 200 words per minute, according to [11]). This discrepancy points to the fact that credibility assessment is a longer and more complicated process than reading comprehension. In case where an expert has many sentences to evaluate, factors such as monotony and monotone may influence time of the assessment as well.

In the current study we used standardized length factor as a measure to approximate the amount of time and attention needed to evaluate a sentence. It is however possible that other factors might affect the final time as well, such as complexity of the extracted text, difficulty with relating tags to the actual content or connecting pronouns with suggested nouns and adjectives. We suspect that some methods might be more time-consuming than the others, considering different types of thinking they impose on a reader. A detailed study of time required to evaluate the text obtained by the presented methods is yet to be performed and should be addressed in the future work.

We therefore propose a simplification, which allows to estimate a proportional shortening of the expert's time in relation to the situation in which he or she is forced to get acquainted with the full context before assessing the sentence. Basing on our LF measure, our best method $TextRank + RB$ allows the expert to assess the credibility of the sentence approximately in 4% of the time needed

for credibility assessment of a full article, and in 25% of the time needed for assessment using our baseline Context Window method.

Based on our experience from the previous study [14] that showed the overall consistency of the credible and non-credible articles (credible articles contain mostly credible sentences and vice versa, ordering of credible/non-credible sentences is not important), we were not surprised by these results. We can hypothesize that non-credible medical content is built upon some finite set of key phrases and words that can be easily detected by an expert. The keyword-based context enrichment methods therefore proved to be suitable to extract the most important features from the context that allow fast and accurate credibility assessment.

6 Conclusions and Future Work

In this article, we have studied the problem of context enrichment for fast fact-checking (credibility evaluation) of sentences. Individual sentences - being part of some larger content - can be evaluated by experts in full context (experts read the entire text to which the sentence belongs) or with partial or no context. We have studied the problem of automatic context enrichment empirically on the case of medical Web content in the Polish language. Our main findings and statements can be summarized as follows:

- In our study the best-performing methods (TextRank + RB, TFIDF + RB) rely on simple NLP tools (such as taggers) and we are confident that our results can be generalized to other languages.
- Evaluation of individual sentences with partial context provided by our methods is faster than evaluation of the same sentences with full context.
- The context enrichment methods presented in this article are not domain-specific and could be applied to fact-checking tasks apart from the medical domain.
- Short textual information acquired in the process of context enrichment could also prove useful in information retrieval tasks.
- As shown by our analysis of types of sentences that cannot be evaluated due to missing context, the problem of context enrichment is a general NLP problem that differs significantly from other NLP problems such as pronoun matching or coreference resolution.

References

1. Buczyński, A., Przepiórkowski, A.: Spejd demo: An open source tool for partial parsing and morphosyntactic disambiguation (2008)
2. Chen, Y.Y., Li, C.M., Liang, J.C., Tsai, C.C.: Health information obtained from the internet and changes in medical decision making: questionnaire development and cross-sectional survey. *J. Med. Internet Res.* **20**(2), e47 (2018)
3. ELRA: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012 (2012)

4. Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W.: Context: an algorithm for determining negation, experienter, and temporal status from clinical reports. *J. Biomed. Inform.* **42**(5), 839–851 (2009)
5. Hill, F., Bordes, A., Chopra, S., Weston, J.: The goldilocks principle: Reading children’s books with explicit memory representations. arXiv preprint [arXiv:1511.02301](https://arxiv.org/abs/1511.02301) (2015)
6. Kadlec, R., Schmid, M., Bajgar, O., Kleindienst, J.: Text understanding with the attention sum reader network. arXiv preprint [arXiv:1603.01547](https://arxiv.org/abs/1603.01547) (2016)
7. Kałol, M., Jankowski-Lorek, M., Abramczuk, K., Wierzbicki, A., Catasta, M.: On the subjectivity and bias of web content credibility evaluations. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1131–1136. ACM (2013)
8. Kałol, M., Nielek, R., Wierzbicki, A.: Understanding and predicting web content credibility using the content credibility corpus. *Inf. Process. Manage.* **53**(5), 1043–1061 (2017)
9. Kopeć, M.: Zero subject detection for Polish. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers, pp. 221–225. Association for Computational Linguistics, Gothenburg, Sweden (2014)
10. Kopeć, M., Ogrodniczuk, M.: Creating a coreference resolution system for Polish. In: LREC [3], pp. 192–195
11. Lewandowski, L.J., Coddling, R.S., Kleinmann, A.E., Tucker, K.L.: Assessment of reading rate in postsecondary students. *J. Psychoeducational Assess.* **21**(2), 134–144 (2003)
12. Logeswaran, L., Lee, H.: An efficient framework for learning sentence representations. arXiv preprint [arXiv:1803.02893](https://arxiv.org/abs/1803.02893) (2018)
13. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (2004)
14. Nabożny, A., Balcerzak, B., Wierzbicki, A.: Automatic credibility assessment of popular medical articles available online. In: Staab, S., Koltsova, O., Ignatov, D.I. (eds.) SocInfo 2018. LNCS, vol. 11186, pp. 215–223. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01159-8_20
15. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Coreference in Polish: Annotation, Resolution and Evaluation. Walter De Gruyter (2015). <http://www.degruyter.com/view/product/428667>
16. Ogrodniczuk, M., Lenart, M.: Web Service integration platform for Polish linguistic resources. In: LREC [3], pp. 1164–1168
17. Przepiórkowski, A., Górski, R.L., Łazinski, M., Pezik, P.: Recent developments in the national corpus of Polish. *NLP, Corpus Linguistics, Corpus Based Grammar Research*, p. 302 (2010)
18. Ren, P., Chen, Z., Ren, Z., Wei, F., Ma, J., de Rijke, M.: Leveraging contextual sentence relations for extractive summarization using a neural attention model. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 95–104. ACM (2017)
19. Samuel, H., Zaïane, O.: MedFact: towards improving veracity of medical information in social media using applied machine learning. In: Bagheri, E., Cheung, J.C.K. (eds.) Canadian AI 2018. LNCS (LNAI), vol. 10832, pp. 108–120. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-89656-4_9
20. National Academies of Sciences, Engineering and Medicine: Improving diagnosis in health care. National Academies Press (2016)

Chapter 5. Articles comprising the thesis

21. Shibuki, H., Nagai, T., Nakano, M., Miyazaki, R., Ishioroshi, M., Mori, T.: A method for automatically generating a mediatory summary to verify credibility of information on the web. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 1140–1148. Association for Computational Linguistics (2010)
22. Turek, P., Wierzbicki, A., Nielek, R., Datta, A.: WikiTeams: how do they achieve success? *IEEE Potentials* **30**(5), 15–20 (2011)
23. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Advances in Neural Information Processing Systems, pp. 2692–2700 (2015)
24. Wang, L., Tu, Z., Way, A., Liu, Q.: Exploiting cross-sentence context for neural machine translation. arXiv preprint [arXiv:1704.04347](https://arxiv.org/abs/1704.04347) (2017)
25. Waszczuk, J.: Harnessing the CRF complexity with domain-specific constraints. the case of morphosyntactic tagging of a highly inflected language. In: Proceedings of COLING 2012, pp. 2789–2804 (2012)
26. Wierzbicki, A.: *Web Content Credibility*. Springer, Heidelberg (2018). <https://doi.org/10.1007/978-3-319-77794-8>
27. Yamamoto, Y.: Disputed sentence suggestion towards credibility-oriented web search. In: Sheng, Q.Z., Wang, G., Jensen, C.S., Xu, G. (eds.) APWeb 2012. LNCS, vol. 7235, pp. 34–45. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29253-8_4

- 5.2 Article 2 "Active Annotation in Evaluating the Credibility of Web-Based Medical Information: Guidelines for Creating Training Data Sets for Machine Learning" (Journal of Medical Internet Research, Medical Informatics, 70 pts.)**

Chapter 5. Articles comprising the thesis

Original Paper

Active Annotation in Evaluating the Credibility of Web-Based Medical Information: Guidelines for Creating Training Data Sets for Machine Learning

Aleksandra Nabożny¹, MSc; Bartłomiej Balcerzak², PhD; Adam Wierzbicki², Prof Dr Hab; Mikołaj Morzy³, PhD, DSc; Małgorzata Chlabicz⁴, MD, PhD

¹Department of Software Engineering, Gdańsk University of Technology, Gdańsk, Poland

²Polish-Japanese Academy of Information Technology, Warsaw, Poland

³Faculty of Computing and Telecommunications, Poznań University of Technology, Poznań, Poland

⁴Department of Population Medicine and Lifestyle Diseases Prevention, Medical University of Białystok, Białystok, Poland

Corresponding Author:

Aleksandra Nabożny, MSc

Department of Software Engineering

Gdańsk University of Technology

11/12 Gabriela Narutowicza St

Gdańsk, 80-233

Poland

Phone: 48 602327778

Email: aleksandra.nabozny@pja.edu.pl

Abstract

Background: The spread of false medical information on the web is rapidly accelerating. Establishing the credibility of web-based medical information has become a pressing necessity. Machine learning offers a solution that, when properly deployed, can be an effective tool in fighting medical misinformation on the web.

Objective: The aim of this study is to present a comprehensive framework for designing and curating machine learning training data sets for web-based medical information credibility assessment. We show how to construct the annotation process. Our main objective is to support researchers from the medical and computer science communities. We offer guidelines on the preparation of data sets for machine learning models that can fight medical misinformation.

Methods: We begin by providing the annotation protocol for medical experts involved in medical sentence credibility evaluation. The protocol is based on a qualitative study of our experimental data. To address the problem of insufficient initial labels, we propose a preprocessing pipeline for the batch of sentences to be assessed. It consists of representation learning, clustering, and reranking. We call this process active annotation.

Results: We collected more than 10,000 annotations of statements related to selected medical subjects (psychiatry, cholesterol, autism, antibiotics, vaccines, steroids, birth methods, and food allergy testing) for less than US \$7000 by employing 9 highly qualified annotators (certified medical professionals), and we release this data set to the general public. We developed an active annotation framework for more efficient annotation of noncredible medical statements. The application of qualitative analysis resulted in a better annotation protocol for our future efforts in data set creation.

Conclusions: The results of the qualitative analysis support our claims of the efficacy of the presented method.

(*JMIR Med Inform* 2021;9(11):e26065) doi: [10.2196/26065](https://doi.org/10.2196/26065)

KEYWORDS

active annotation; credibility; web-based medical information; fake news

Chapter 5. Articles comprising the thesis

Introduction

Background

In 2020 and 2021, the world has not been fighting only a pandemic; more precisely, it has been fighting both a pandemic and an infodemic [1]. The spread of COVID-19 has been accompanied by an equally unfortunate and dangerous spread of misinformation such as fake news linking the COVID-19 epidemic to 5G technology [2]. Disinformation has influenced other disease outbreaks such as the measles outbreak in Germany that involved more than 570 reported measles cases and caused infant deaths [3]. This study suggests that there exist numerous similar examples. From anticholesterol treatment to psychiatry—potentially harmful noncredible medical content on varied topics proliferates on the web.

Web-based information related to health and medicine is a large and influential category of web content, to the extent that the term *Dr Google* has been coined. The case of health-related web content is interesting from the point of view of informatics not only because medical information is highly specialized and written using domain-specific vocabulary, but also because medical information on the web is often misinterpreted or taken out of context. Health-related fake news reports often rely on factually correct medical statements such as the antiseptic effect of silver ions, which translates into a false belief in the universal effectiveness of colloidal silver for treating any disease. Debunking health-related web content requires not only expertise but also awareness of the possible effects of misinterpreted information. The breadth of specialized medical knowledge, coupled with the impact of context on fake news debunking, increases the difficulty of the problem of medical fake news detection.

Fully automated methods are currently not mature enough to detect medical fake news with sufficient accuracy. A realistic system for detecting and debunking medical fake news needs to keep medical experts in the loop. However, such an approach is not scalable because medical experts and health professionals cannot allocate sufficient time to handle the volume of misinformation spreading on the web. Another issue is that, in general, compared with credible medical content, noncredible medical web content is sparse. Assuming a real human-assisted system for assessing the credibility of medical statements, statistically, out of 100 assessed statements, the expert will catch no more than 20 unreliable items (as shown by our data collection experiment). The purpose of our work is to create an automatic tool to maximize the number of potentially noncredible sentences to be verified in the first place. The sentences are then reordered so that the most noncredible content shows up first to be annotated by a human judge. In such a way, we can optimize medical experts' time and efficacy when annotating medical information. Even if only a portion of potentially noncredible sentences gets annotated by the expert, it will include the most suspicious content.

We propose to use a method called active annotation. It dramatically improves the use of annotators' time. Active annotation implements a highly efficient human-in-the-loop component for augmented text annotation. The main idea behind

active annotation is to use an unsupervised machine learning method (grouping of sentences into clusters based on sentence similarity) to organize the training data to suggest annotation labels for human annotators. When active annotation is used, the work of human annotators (medical experts) is focused on difficult noncredible medical statements. In addition, because the annotators process clusters of semantically similar sentences, our method significantly reduces the cost of cognitively expensive context switching. However, it is the annotators who decide the final labeling of the data.

The method proposed in this paper extends currently known active annotation methods by a cluster-ranking procedure that ensures that medical experts first see the content clusters that are most likely to contain noncredible content. This approach allows us to speed up the discovery of noncredible content. In our view, the process of detecting and debunking medical misinformation will never stop, and therefore a method that optimizes the use of medical experts' is of essential importance.

To test our method, we conducted an experiment with the participation of medical experts. They were asked to evaluate the credibility of medical and health-related Web content. The result of the experiment is a large data set that contains numerous examples of medical misinformation. We conducted an explorative and qualitative analysis of this data set, searching for patterns of similarity among the different examples of medical misinformation. The result of this analysis (which included an in-depth case study of misinformation related to cholesterol therapy with statins) was the discovery of distinct narratives of medical misinformation. We believe that these narratives are general in nature and will be of great use for detecting medical misinformation in the future.

Our direct experiences with the annotation team dictate a set of rules that have been formalized as a strict protocol for medical text annotation. Most importantly, we noted that the annotators tended to use external contexts extensively when annotating data. This, in turn, led to incoherent annotation labels across the data set and a divergence between the notions of statement credibility and statement truthfulness. We share our experience and present an annotation protocol that we have used to mitigate some of the annotation problems.

The original contributions presented in our paper are as follows:

- An annotation schema, an annotation protocol, and a unique annotated data set comprising 10,000 sentences taken from web-based content on medical issues, labeled by medical experts as credible, noncredible, or neutral. The entire data set is available in a public repository [4].
- A method for ranking sentences submitted to medical experts for labeling. Our active annotation method increases the likelihood that medical experts will discover noncredible sentences and thus optimizes the use of medical experts' time.
- A qualitative analysis of the labeled data set. We discovered 4 distinct narratives (both syntactic and semantic) present in the noncredible statements. We believe that these narratives can be further used to discern noncredible statements in areas of medicine other than the areas covered by our data set.

Chapter 5. Articles comprising the thesis

Literature Review

Health literacy is a rising concern, especially during the COVID-19 pandemic. However, research shows that more than half of the population struggles with making proper judgments and taking decisions in everyday life concerning their health [5]. Moreover, studies from the United States, Europe, and Australia [6,7] have found that web-based health information is written above the average reading level of adults. There is clearly the need for external tools or strategies to support laypersons in assessing the credibility of web-based health information. Expert fact-checking is one of the proposed strategies [8] because short-format refutational medical expert fact-checks have proven to be free from the *backfire effect* [9] (the *backfire effect* has been described in the study by Nyhan and Reifler [10]). Research shows that using expert sources to correct health misinformation in social media permanently corrects users' false beliefs.

The related work on the general news media domain [11] demonstrates that a credible source can promote false information and vice versa. Technological innovation in the fight against disinformation, as the authors argue, should go beyond discrediting noncredible sources of information and should instead promote more careful information consumption [11]. The literature has reported on successful machine learning models that classify entire articles or information sources [12,13]. Of note, these models can easily overfit (ie, obtain high classification accuracy for publications from media outlets present in the training set but fail to generalize to previously unseen media outlets). The possible performance drop in classifying fake news from previously unseen sources has been examined in the literature [12]. The study by Afsana et al [14] is, to the best of our knowledge, the most accurate classification model for assessing the quality of web-based health information. The authors declare accuracy ranging from 84% to 90% varied over 10 criteria. The model includes source-level and article-level features. The relationship of the described criteria with credibility remains an open research question.

The assessment of the veracity of individual claims contained in open-domain news articles is an emerging and fast-growing field of research. The scope of activities includes the creation of data sets containing the claims collected from fact-checking websites, such as MultiFC [15], Liar [16], and Truth of Varying Shades [17], and the existing solutions are based on a variety of approaches, from semi-automatic knowledge graph creation [18] to choosing check-worthy claims and comparing them against verified content (ClaimBuster) [19]. The open-domain solutions or solutions used in journalism [20] are not easily transferable to the medical domain.

The MedFact system [21] is a stand-alone web-based health information consumption support system. In MedFact, the user is automatically equipped with relevant trusted sources during web-based discussions.

State-of-the-art information retrieval models [22,23] forms part of the fully and semi-automatic fact-checking systems. A combination of such systems' judgments and human judgments has been successfully applied in the study by Ghenai and Mejova [24] for the specific case of capturing the spread of rumors

regarding the Zika virus. Our goal is to test the combination of an unsupervised machine learning model with a human-in-the-loop approach as a robust tool to support the assessment of the credibility of web-based medical statements.

The quality assessment coding scheme for lay medical articles had been proposed in the 1990s under the Discern handbook project [25] and as Health on the Net (HON) principles. However, the guidelines have to comply with the rapidly evolving web-based reality; thus, new tools and updates are designed every few years, such as the Ensuring Quality Information for Patients (2004) [26] tool, Evidence-Based Patient Information (2010) [27], and Good Practice Guidelines for Health Information (2016) [28], to name a few. Keselman et al [29] propose different credibility assessment criteria based on 25 web-based articles regarding type 2 diabetes. These criteria (objectivity, emotional appeal, promises, and certainty) can be automatically captured by language models and lexicon-based machine learning. Work on web-based journalism has developed good practices that can also be used by medical experts in credibility evaluation. Medical practitioners who directly communicate medical information to patients can observe their reactions and subsequent actions and therefore have a special agency in credibility evaluation.

Successful application of machine learning models requires the annotation of vast corpora of medical information. However, this annotation is prohibitively expensive given the required expertise of the annotators and their limited capacity. Active annotation is a technique that facilitates large-scale data annotation by providing an auxiliary ranking of sentences that should be manually annotated by medical experts and by expediting labeling of other sentences to the underlying machine learning model. In this study, we are particularly inspired by the approach presented by Marinelli et al [30]. The authors propose initially dividing text documents into separate clusters, selecting pivot documents (k-closest documents to the center of each cluster), and generating a tentative label for the cluster. Next, a small set of text documents is selected and presented to human annotators with a proposed label and a binary annotation decision (to accept or reject the label). The authors claim that in many applications, obtaining a full annotation schema before annotation may be difficult and turning the annotation task into a binary question-answering task significantly speeds up the process [30].

Language Modeling

The term *language model* is confusing because it serves as an umbrella term for different concepts. As a general rule, a language model is a way in which textual content (tokens, words, sentences, paragraphs, and documents) is represented. Historically, text documents have been represented using 2 prevalent models: the bag-of-words model (where a document is represented simply as the set of words appearing in the document) and the one-hot encoding model (where a document is represented by a binary vector of a length equal to the size of the vocabulary and each position in the vector encodes the presence or absence of a word in the document). The most consequential limitation of these models was the inability to capture the semantic similarity between words. For instance, if

Chapter 5. Articles comprising the thesis

a document contained the word *diabetes* and another document contained the word *insulin*, there was no straightforward way of deciding that the documents shared a common topic. This limitation has been abruptly neutralized with the advent of word embeddings. Word embeddings are dense continuous vector representations of words from a given vocabulary, which means that each word is assigned a unique vector whose elements are arbitrary numbers. Unlike one-hot encoding vectors where each vector has a length equal to the size of the vocabulary, word embedding vectors have, at most, several hundred dimensions. The vectors are trained on the text corpus to capture various semantic relationships among words. For instance, words such as *apple*, *pear*, and *orange* appear close to each other in the vector space because part of their representation encodes the notion of being a fruit. Analogically, the distance between the words *Russia* and *Moscow* is similar to the distance between the words *Great Britain* and *London* because the difference between the respective word vectors encodes the notion of a capital city.

Since the seminal work of Mikolov et al [31], word embeddings have revolutionized the field of natural language processing. After the initial success of the *word2vec* algorithm, numerous alternatives have been introduced: Global Vector embeddings trained through matrix factorization [32], embeddings trained on sentence dependency parse trees [33], embeddings in the hyperbolic space [34], subword embeddings [35], and many more. The common feature of these embeddings is the static assignment of dense vector representations to words. Each word receives the same embedding vector, irrespective of the context in which the word appears in a sentence. These static embeddings can be used to create representations for larger text units such as sentences, paragraphs, and documents. However, static embeddings are inherently unable to capture the intricacies hidden in the structure of the language and encoded in the context in which each word appears. Consider these 2 sentences: “A photo reveals significant damage to the tissue” and “Please do not throw used tissues into the toilet.” The word *tissue* will receive the same vector although the context allows disambiguation of the meaning of the word.

To mitigate this limitation, modern language models depend on deep neural network architectures to calculate accurate, context-dependent word and sentence embeddings. First, context-dependent language models used either the long short-term memory network architecture [36] or gated recurrent unit networks [37] to capture contextual dependencies among the words appearing in a sentence. In other words, unlike static word embeddings, context-dependent language models calculate an embedding word vector based on the context (ie, words surrounding the embedded words). In the aforementioned example, the word *tissue* would receive 2 different vector representations: in the first sentence, the vector for the word *tissue* would be much closer to the vectors of words such as *skin* or *cell*; in the second sentence, the vector for the word *tissue* would be closer to the vector of the word *handkerchief*. These early recurrent architectures, however, suffered from performance drawbacks, and in 2018 they were replaced by transformer architecture [38]. This architecture allowed the training of much better embeddings, such as Google’s Universal

Sentence Encoder [39] or the (infamous) Generative Pre-trained Transformer 3 [40].

The current state-of-the-art language model, Bidirectional Encoder Representations from Transformers (BERT) [41], produces continuous word vector representations by training the neural network using 2 parallel objectives: guessing the masked word in a sentence (ie, trying to predict the word based on the context) and deciding whether 2 sentences appear one after another. Given such training objectives, the network applies similar weights to the nodes regarding input words that appear in a similar context. Sentence-BERT (sBERT) [42] is a straightforward extension of the original BERT architecture for creating sentence embeddings. This model is based on Siamese BERT networks [43] (2 identical models trained simultaneously) that are fine-tuned on the Natural Language Inference and Semantic Textual Similarity tasks. The model serves as an encoder for sentences. The encoder calculates vector representations of sentences so that semantically similar sentences have low cosine distance in the latent embedding space. This is both more efficient and produces semantically richer sentence representations than simply averaging the vectors of words that appear in each sentence.

Methods

Presentation of 3 Steps

To validate the efficacy of the active annotation approach, we need to create a data set of sentences on medical topics gathered from the Web, after which we need to obtain credibility evaluations of these sentences from medical experts. We need to propose methods for selecting sentences from the Web, annotating of these sentences by medical experts, and organizing these sentences into a processing pipeline to use the experts’ time and attention most efficiently. These 3 steps we elaborate on in this section.

Data Selection

We performed annotation on a data set of 247 articles collected manually from various eHealth websites. The data set consists of more than 10,000 sentences. All documents were annotated by medical professionals sentence by sentence. The sentences constitute a stratified sample of source texts of varying credibility. We first discussed the most problematic topics of specific medical fields with the medical practitioners. Next, we manually searched for articles that presented contradicting views regarding these topics. These topics include the following:

1. Pediatrics:
 - Children’s antibiotics consumption (432 sentences)
 - Children’s steroids consumption (701 sentences)
 - Vaccination (1262 sentences)
 - Dietary interventions for children with autism (431 sentences)
 - Food allergy testing (1401 sentences)
2. Psychiatry:
 - Effectiveness of psychiatric medication and electroconvulsive therapy (2272 sentences)
3. Cardiology:

Chapter 5. Articles comprising the thesis

- Benefits of statin therapy in treating cardiovascular disease (CVD; 2029 sentences)
 - Dietary interventions for heart health improvement (423 sentences)
 - Benefits of consumption of antioxidants (694 sentences)
4. Gynecology:
- Benefits of cesarean section over natural birth (359 sentences)
 - Selective serotonin reuptake inhibitor consumption during pregnancy (169 sentences)
 - Aspirin consumption during pregnancy (257 sentences)

Our collection of web-based health-related and medical articles reflects topics potentially causing controversy and misinformation among patients.

Methodology of Selecting Source Websites

The source websites were selected as follows. First, we asked each medical practitioner 2 questions:

1. “In your medical practice, what kind of false beliefs and rumors do you encounter when interacting with patients?”
2. “The truthfulness of which facts do you have to prove to your patients most often?”

The answers to these questions served as the basis for manually creating web queries. To create a data set of web medical articles addressed to laypersons, we submitted these queries to the Google search engine and then manually selected sources. The full list of these queries is listed in [Multimedia Appendix 1](#). The manual collection was supported by the HON browser plugin (HON tag-certified webpages). As a result, 12.6% (31/247) of the extracted articles originated from HON-certified sources. The remaining 87.4% (216/247) come from domains such as the following:

- Large news media outlets (eg, *The Guardian*, *The New York Times*, and BBC)
- Q&A forums, both general and topic-specific (eg, “Quora”, “Yahoo”, “community.babycenter.com”)
- Parenting blogs (eg, “scarymommy.com”)
- Uncertified health portals (eg, “choosingwisely.org”, “practo.com”, and “heartuk.org.uk”)
- Advertising websites for medical supplements and medical testing (eg, “everlywell.com”, “yorktest.com”, and “naturesbest.co.uk/antioxidants”)

The full list of data sources is available in [Multimedia Appendix 2](#).

In this study, we consider a sentence as the unit of consistent information that undergoes credibility assessment. According to Wikipedia [44], “a sentence is a set of words that in principle tells a complete thought.” Thus, unless a sentence is highly complex, we can assume that the segmentation of a document into sentences is the easiest way to automatically extract single statements. To be precise, a single sentence may contain several statements. We have also observed that expert annotators tend to focus on statements rather than entire sentences when labeling data. However, we do not have a robust method of statement demarcation. In addition, most sentences contain a single main

statement; thus, we decided to make the sentence the atomic unit of annotation and classification.

An additional reason for focusing on single sentences is the phenomenon of shrinking attention. Recent studies suggest that, over recent decades, collective attention spans are becoming shorter across all domains of culture, including the web [45]. It is debatable as to what the underlying cause of this phenomenon is. The most likely explanations suggest the impact of the rapid acceleration in the rate of production and consumption of information. Given finite attention resources, this inevitably leads to more cursory interaction with information. It is possible that this phenomenon also affects the consumption of health-related information, which only exacerbates the problem of the ubiquitousness of medical fake news on the web.

Expert Annotators

In all, 9 medical professionals took part in the experiment: 2 cardiologists, 1 gynecologist, 3 psychiatrists, and 3 pediatricians. All the experts had completed 6 years of medical studies, followed by a 5-year specialization program that culminated in a specialization examination. The experts were paid for a full day of work (approximately 8 hours each). Of the 9 experts, 8 (89%) had at least 10 years of clinical experience. The gynecologist was a resident physician; we accepted his participation in the experiment because of his status as a PhD candidate in medicine. Of the 3 psychiatrists, 1 (33%) held a PhD degree in medical sciences. The experts were allowed to browse certified medical information databases throughout the experiment. Each expert evaluated the credibility of content within their specialization (cardiology, gynecology, psychiatry, or pediatrics).

Annotation Protocol

Our goal is to create a rich and diverse corpus of medical sentences assessed and labeled in terms of their credibility by medical experts. To obtain reliable and comparable credibility evaluations, the experts participating in our study were supported by a detailed annotation protocol.

The medical experts evaluated the credibility of sentences with the following set of labels and the corresponding instruction:

- CRED (credible): the sentence is reliable; does not raise major objections; contains verifiable information from the medical domain
- NONCRED (not credible): the sentence contains false or unverifiable information; contains persuasion contrary to current medical recommendations; contains outdated information
- NEU (neutral): the sentence does not contain factual information (eg, it is a question); is not related to medicine

The experts were asked to base their answers mostly on their experience, knowledge, and intuition, but they were also allowed to use an external database that they would usually use in the course of their medical practice. The main direction provided to the experts was to focus on the patient’s alleged perception of the information. The control question stated as follows: “If

Chapter 5. Articles comprising the thesis

the patient asked you if he or she should trust this statement, would you say yes or no?"

In addition, we collected the following information for each sentence:

- Time needed for evaluation (in milliseconds)
- (Optional) Reason for evaluating the sentence as noncredible
- Number of surrounding sentences needed to understand the context of the sentence being evaluated

Examples of credible sentences from the *cholesterol and statins* topic include the following:

Lp(a), the worst cholesterol, is a number most doctors don't measure.

Monitoring cholesterol levels is crucial because individuals with unhealthy cholesterol levels typically do not develop specific symptoms.

Non-communicable chronic disease is now the biggest killer on the planet.

Examples of noncredible sentences include the following:

For the remaining 90% of the population, the total cholesterol had no predictive value.

It seems likely that fear of fat is unreal, based on a carry-on of the cholesterol fear.

Most people don't need to cut down on the cholesterol that's found in these foods.

Examples of neutral sentences include the following:

Seven [research items] found no link between LDL cholesterol and cardiovascular mortality.

These perspectives won't make headlines and they won't appeal to those who want a simple and definite answers.

This is not why I went to medical school.

Impact of Sentence Context on Credibility Evaluation

Table 1 shows how many sentences required additional m -surrounding sentences to provide the context for annotation. When focusing on noncredible statements, more than 71.27% (1377/1932) of the sentences were self-explanatory, 26.6% (514/1932) of the sentences required a single sentence of context, and less than 2.17% (42/1932) of the sentences required 2 or more sentences of context. Thus, we conclude that our choice of the sentence as the unit of information is justified.

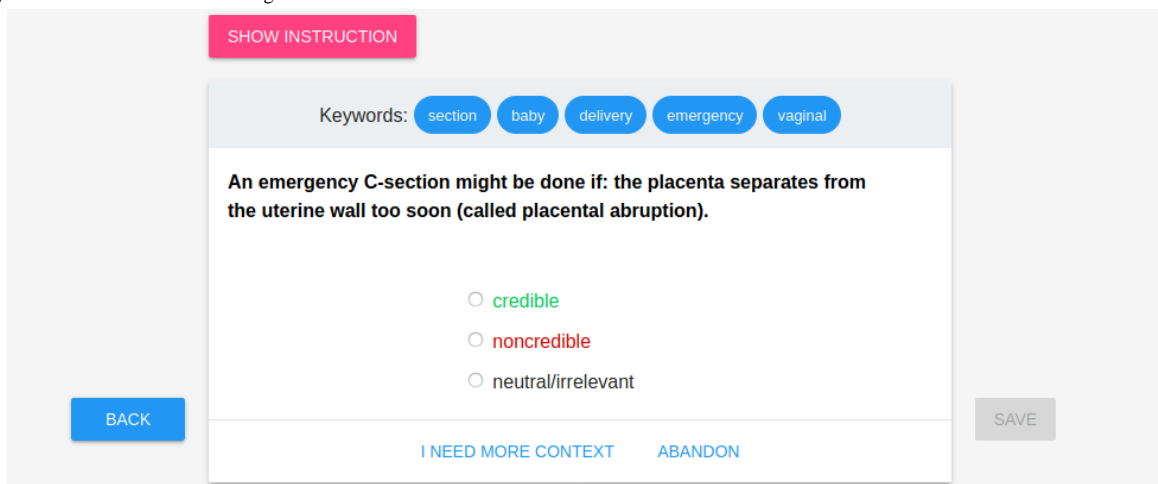
Table 1. Number of surrounding sentences (m) needed to understand the context and evaluate the credibility of a sentence for all data, only credible subset, only noncredible subset, and only neutral subset ($n=10,649$).

m	All data, n (%)	Credible subset, n (%)	Noncredible subset, n (%)	Neutral subset, n (%)
0	8565 (80.43)	4955 (80.07)	1377 (71.27)	2233 (88.3)
1	1958 (18.39)	1165 (18.83)	514 (26.6)	279 (11.03)
2	107 (1)	57 (0.92)	34 (1.76)	16 (0.63)
3	12 (0.11)	5 (0.08)	6 (0.31)	1 (0.04)
<3	8 (0.07)	6 (0.1)	2 (0.05)	0 (0)

For the annotation process, we used the software developed specifically for this experiment. During the experiment, the medical expert could not see the context of the whole document while annotating a sentence. However, we provided the most

relevant keywords collected from the rest of the document. Keywords were extracted using the methods described in the study by Nabožny et al [46]. A single task is shown in Figure 1.

Figure 1. Annotation interface: single sentence view.



Chapter 5. Articles comprising the thesis

If the medical expert decided that a sentence could not be assessed because of insufficient context (despite visible keywords), they could display the preceding and succeeding sentences in the annotation view, as shown in Figure 2. Each medical expert was asked to annotate approximately 1000 randomly chosen sentences. Whenever the medical expert

labeled a sentence as noncredible, they were asked to provide the reason for their decision. To avoid the effect of intentionally skipping the NONCRED label to complete the task quicker, providing the reason was optional, and the expert could also choose an explanation from a set of tags prepared beforehand.

Figure 2. Annotation interface: sentence in context view.

SHOW INSTRUCTION

Meanwhile, all researchers believe that any woman who wants to get pregnant should take aspirin, unless she is allergic to this drug and does not suffer from gastric problems.

However, opinions on this topic are divided, because many experts are afraid, in turn, that acetylsalicylic acid may promote bleeding, which pose a serious threat to the fetus.

Before you start taking medicine, consult your doctor!

credible

noncredible

neutral/irrelevant

BACK

I NEED MORE CONTEXT ABANDON

SAVE

The set of possible explanations prepared in advance included the following:

- The sentence contains argumentation that is weak or irrelevant, given the context of the subject being discussed.
- The sentence contains an encouragement to act inconsistently with current medical knowledge.
- The author of this sentence shows signs of the lack of substantive knowledge or is not objective.
- The sentence is an anecdote or a rumor.
- The sentence is an advertisement of an unproven drug or substance or an unproven therapy.
- The sentence cites research that was conducted on a small sample.
- The sentence contains invalid numerical data.
- The sentence contains outdated information.
- The sentence is incomprehensible or grammatically incorrect.

Most of the annotation was conducted in controlled laboratory conditions. The experts were performing annotation tasks in the presence of a supervisor who was conducting the experiment. At any time, the medical experts had access to the detailed instruction (definitions of each label) and could also ask the supervisor for assistance. The experts completed 70% of the tasks in controlled conditions, and the rest were completed with web-based assistance within a few days after the conclusion of the laboratory experiment.

Sentence Processing Pipeline Using Clustering and Reranking

Inspired by the active learning paradigm, we designed an assessment loop for medical sentence credibility. The core idea of the active annotation approach is to augment annotation efforts by 2 mechanisms:

- *Clustering:*
Semantically similar sentences are automatically grouped into clusters. The process of clustering uses sentence-embedding representation. Each sentence is represented as a vector computed by the language model. As each sentence is a vector, mathematical measures of a distance can be used, such as the Euclidean distance or the cosine distance. We use the k-means algorithm to divide sentences into clusters. K-means is a simple iterative procedure where clustered items (in our case, vectors representing sentences) are assigned to the closest of k points representing cluster centers (also known as centroids). After assigning each item to the nearest centroid, the positions of the centroids are updated to reflect the geometric mean of assigned items. Finally, items are reassigned to the nearest centroid, and the procedure is repeated until no more reassignments are possible. The resulting clustering maximizes the similarity among the items assigned to a cluster and at the same time minimizes the similarity among the items assigned to different clusters. In other words, if 2 sentences are assigned to the same cluster, the distance between their vector representations is small, which in turn means that the sentences are

Chapter 5. Articles comprising the thesis

semantically similar (because semantic similarity is the criterion of embedding vector training). When human annotators are presented with sentences from a cluster, they process sentences that share a common topic. This reduces the cognitive workload of human annotators because they do not have to switch contexts between annotated sentences.

- **Reranking:**

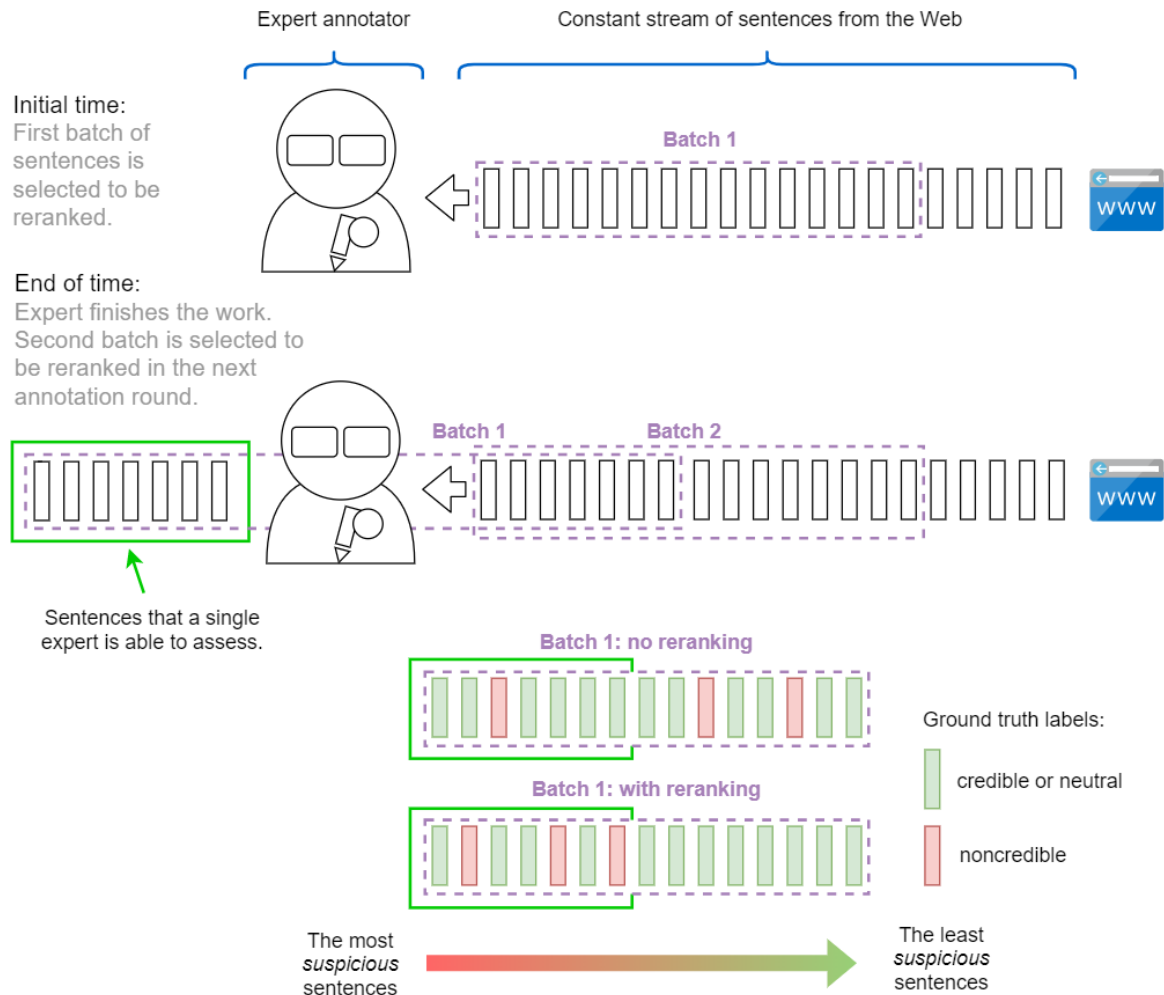
Noncredible statements are moved to the top of the ranking. Human annotators are required to identify noncredible statements; thus, every time human annotators are presented with a credible or neutral sentence, they may consider it to be a waste of their precious time. By combining sentence embeddings and clustering, we push sentences that are close to the already labeled noncredible sentences to the top of the ranking, prioritizing these sentences for the next round of manual annotation.

In the active annotation process, the following steps are performed in the assessment loop:

1. Sentences from the corpus are encoded by the language model to produce sentence embeddings.
2. The k-means clustering algorithm [47] is applied, and the top k sentences nearest to the cluster center are chosen for initial human annotation. We use the elbow method [48] to find the number of clusters (which represents the number of distinct topics in the corpus).
3. Medical experts annotate selected sentences.
4. The algorithm reranks all sentences based on the distribution of labels within clusters.
5. Medical experts annotate sentences from the top of the ranking, triggering another reranking procedure.

The general idea behind reranking is presented in Figure 3.

Figure 3. Sentence reranking: general idea.



Step 4 is crucial to the method. First, we find clusters with a large proportion of labeled noncredible statements. During initial iterations of the method, only a small fraction of sentences are manually labeled, but the clustering step groups semantically similar sentences; therefore, we expect that many sentences belonging to a cluster with predominantly noncredible labels

also would turn out to be noncredible. In step 5, more sentences are manually labeled, providing a better approximation of the true distribution of labels within clusters. By repeating steps 4 and 5, we annotate more and more sentences, prioritizing the annotation of noncredible sentences.

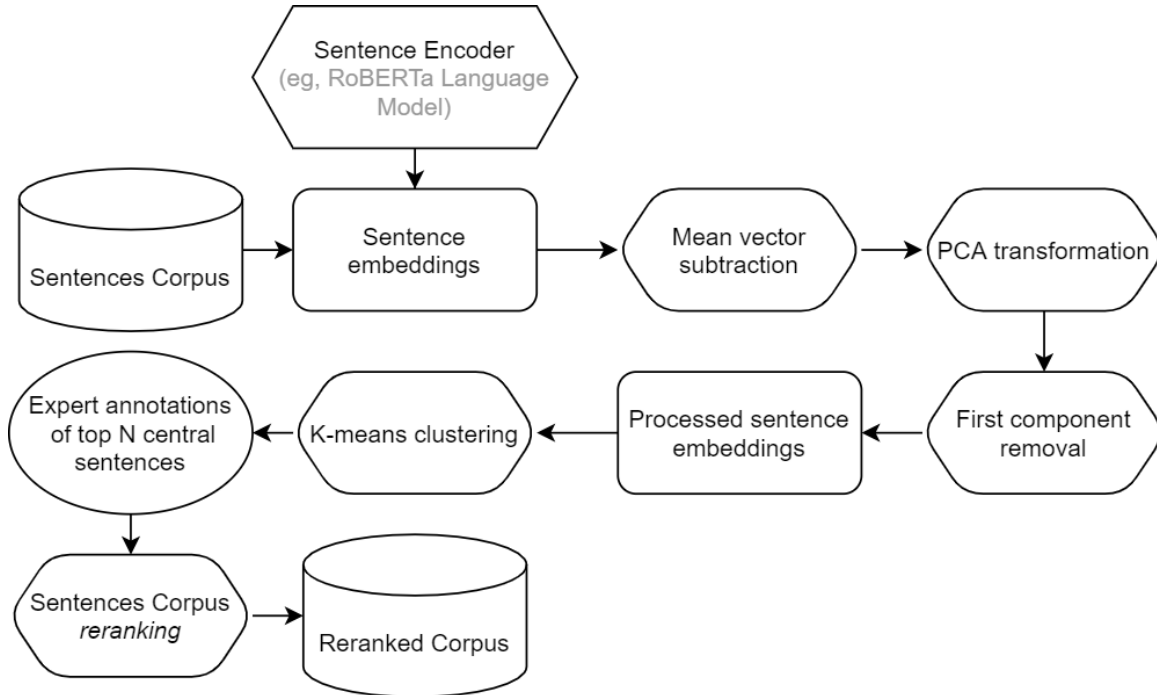
Chapter 5. Articles comprising the thesis

For sentence embeddings computations, we use the sBERT modification Robustly Optimized BERT Pretraining Approach where embeddings are calculated based on the same model as BERT but with slightly different training objectives and hyperparameters [49]. We also use a simple preprocessing technique where we subtract the mean and exclude the first principal component from each embedding vector [50,51] (principal component analysis transformation). The assumption

behind this step is that the first principal component encodes syntactic rules of the grammar of the sentences without contributing to their semantics. The removal of the first component strips sentence vectors of grammar and leaves only the part of the vector where the meaning is encoded.

Figure 4 presents the overview of the sentence processing pipeline.

Figure 4. Processing pipeline. PCA: principal component analysis; RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.



The key component of the pipeline is the clustering and reranking strategy. For reranking, we perform 2-level sorting. The first sorting is applied to clusters, and the second sorting reorders sentences within clusters. We rank clusters based on the proportions of credible, noncredible, and neutral labels in the top m most central sentences. Our scoring formula penalizes clusters with a significant proportion of credible sentences. At the same time, it rewards clusters with a significant proportion of noncredible sentences. This strategy enables us to push most of the noncredible sentences to the top of the ranking, thus positioning them at the top of the queue for medical expert evaluation.

Let $p(c)$, $p(n)$, and $p(u)$ denote the probability that a random sentence is credible, noncredible, or neutral, respectively. This probability is computed by manually annotating m most central sentences in the cluster. The cluster score is defined as follows:

$$\text{score}@k = 1/e^{-(p[n]-p[c])} + 1/w^{p(u)+1}(I)$$

The first component of the formula is the sigmoid function with the difference between $p(n)$ and $p(c)$ as the argument. If the difference is positive, which means that there is an advantage of noncredible proportion over credible, the sigmoid function gives results close to 1 (the bigger the difference, the closer to 1). If the difference is negative, the sigmoid value tends toward

zero. The second component of the formula is the parametrizable function, which enables giving proper scoring weight to $p(u)$. For example, given $w=1.5$, it orders clusters with $p(n)=0.4$ and $p(c)=0.3$ below clusters with $p(n)=0.5$ and $p(c)=0.4$. Without the second component, both clusters would receive the same score.

The intracluster ranking of sentences is performed based on the distance of sentences from the center of the cluster, with more central sentences placed at the top of the ranking. The distance is measured as the cosine distance in the latent embedding space. The final ranking of all sentences is obtained by first ordering all clusters in the decreasing order of $\text{score}@k$ and, next, by reordering sentences within each cluster by the growing distance from the center of the cluster.

Results

Overview

We used the method described in the previous section to create an annotated data set. We now describe the results. First, we present the data set statistics. Next, we depict the effect of our sentence pipelining method on the effectiveness of the medical experts' time allocation. Subsequently, we conduct a qualitative

Chapter 5. Articles comprising the thesis

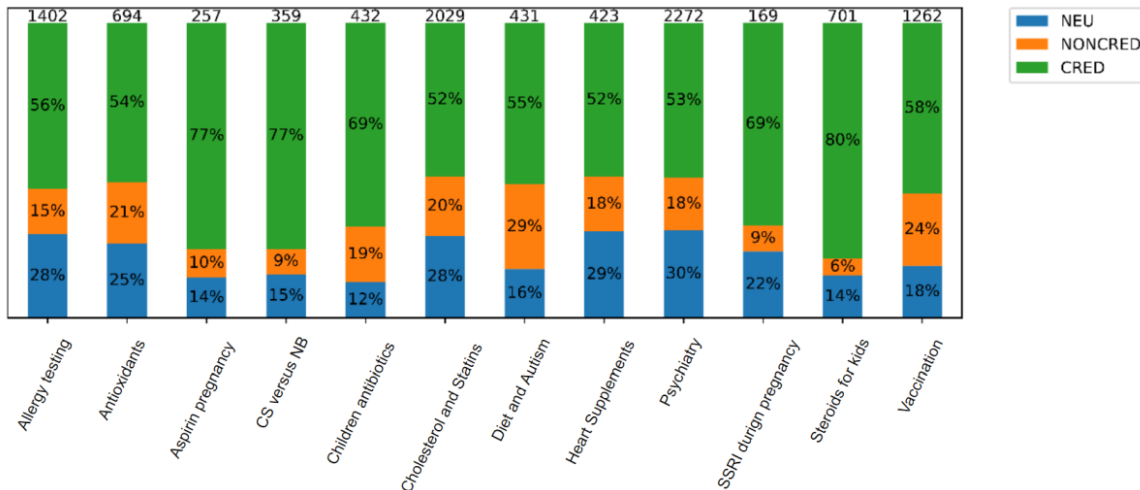
analysis of the credible and noncredible sentences, focusing on a single topic.

Distribution of Labels Within the Data Set

The distribution of labels (CRED, NONCRED, and NEU) for each topic is shown in Figure 5. Distribution varies for each topic but within a certain range. For example, the CRED label

is always at least two times more frequent than the NONCRED label and significantly more frequent than the NEU label. The NEU label applies to no more than 30% (3195/10,649) of the sentences in all topics, which leads us to the conclusion that, regardless of the topic, more than 59.99% (6389/10,649) of the statements warrant credibility checking.

Figure 5. Distribution of credible, noncredible, and neutral sentence labels within topics. CS: cesarean section; CRED: credible; NB: natural birth; NEU: neutral; NONCRED: noncredible; SSRI: selective serotonin reuptake inhibitor.



Although the articles were explicitly picked so that they reflect potentially controversial topics, the proportion of noncredible sentences was generally small. Taking into account the alarm-raising calls of the medical experts, we can conclude that even a small contribution of noncredible content throughout the web has a substantial influence on the formation of people's views.

Justification for Using the Lift Measure

We have chosen the lift measure to evaluate the effectiveness of our method. Throughout the qualitative analysis, it became apparent that semantic similarity measures retrieved from neural language models lose important information encoded in annotations. Our objective is to optimize medical experts' time by focusing their attention on statements that are possibly noncredible. Using the lift measure, we determined the relative time savings by indicating how many more noncredible sentences a medical expert would see by reviewing a given percentage of the entire sentence corpus using our ranking. The lift measure specified for each ranking percentile is defined as follows:

$$lift@p = N/p \times recall@p(2)$$

where p is the percentile, N is the total number of sentences in the corpus, and $recall@p$ defines, for a given percentile p of the ranking, how many noncredible statements have been included in the p th percentile of the ranking.

The key parameter of our method is m , the number of top sentences in a cluster for manual annotation. We tested our method on a full data set (all topics merged) for 3 m values, each of which is listed in Table 2. In Table 3, we present the lift results for the separate topic of *cholesterol and statins*. The baseline value for lift is 1. Thus, we can interpret the results as follows: the number by which a given value exceeds 1 tells us how many more noncredible sentences medical experts would discover at a given corpus percentile when using the reranking procedure. For example, when reviewing 20% of the full corpus, medical experts would discover 29% more noncredible sentences if the batch were to be reranked using the m value of 5 than without applying the procedure.

Table 2. Lift results for the full data set. m is the number of top sentences from each cluster to be manually reviewed.

lift@ m	Number of clusters	Batch percentile			
		1% (approximately 100 sentences)	10% (approximately 1000 sentences)	20%	40%
lift@5	200	1.36	<i>1.36</i> ^a	1.29	1.17
lift@10	130	1.23	1.31	1.3	1.17
lift@15	100	1.49	1.27	1.22	1.16

^aThe best performing set of parameters for a given batch percentile is italicized.

Chapter 5. Articles comprising the thesis

Table 3. Lift results for the cholesterol and statins topic. m is the number of top sentences from each cluster to be manually reviewed.

lift@ m	Number of clusters	Batch percentile			
		1% (approximately 20 sentences)	10% (approximately 200 sentences)	20%	40%
lift@5	40	1.75	1.24	1.26	1.27

The number of clusters for each experiment is chosen based on 2 criteria: the elbow method [48] and the proportion of sentences to be manually reviewed. The latter should not exceed 15% of the batch. Let us take Table 3 as an example: we delegate $5 \times 40 = 200$ top sentences from each cluster to be manually reviewed by the experts. These 200 sentences out of the approximately 2000 sentences in the *cholesterol and statins* topical category make up 10% of the set. It means that by gathering initial labels from only 10% of the sentences from the topical corpus, we can obtain significant (eg, 27% in the 40th percentile) savings of experts' time during text annotation sessions.

Zooming in on a Topical Cluster: Case Study of Statins

We conducted a case study in the subdomain of cholesterol and statins. We did this to gain insight into the process of credibility evaluation and the nature of noncredible medical sentences. The focus on a single topic was dictated by the size and diversity of our data set. Presenting an in-depth qualitative analysis of the entire data set would take too much space. The following is a qualitative analysis of all sentences labeled noncredible by the experts in the selected topic.

Brief Introduction to the Topic of Statin Use

Numerous epidemiological studies, Mendelian randomization studies, and randomized controlled trials have consistently demonstrated a relationship between the absolute changes in plasma low-density lipoprotein (LDL) and the risk of atheromatous CVD. The inverse association between plasma high-density lipoprotein and the risk of CVD is among the most consistent and reproducible associations in observational epidemiology. Higher plasma Lp(a) concentrations are associated with an increased risk of CVD, but it appears to be a much weaker risk factor for most people than LDL cholesterol [52]. Commonly, plasma cholesterol is used to calculate cardiovascular risk, whereas LDL is used to evaluate the achieving of target values according to the estimated cardiovascular risk.

Hypercholesterolemia (dyslipidemia with an increased levels of circulating cholesterol) is not the only factor responsible for the development of CVD, but also obesity, poor diet, lack of physical activity, smoking, and high blood pressure (hypertension). To prevent CVD, physicians recommend that patients quit smoking; eat a diet in which approximately 30% of the calories come from fat, choosing polyunsaturated fats and avoiding saturated fats and trans fats; reduce high blood pressure; increase physical activity; and maintain their weight within normal limits [53].

Hydroxymethylglutaryl-coenzyme A reductase inhibitors (statins) lower cholesterol synthesis. Statins represent the cornerstone for the treatment of hypercholesterolemia and in the prevention of CVD, although muscle-related side effects have strongly limited patients' adherence and compliance [53].

The evidence in support of muscle pain caused by statins is in some cases equivocal and not particularly strong. The reported symptoms are difficult to quantify and rarely is it possible to establish a causal link between statins and muscle pain. In randomized controlled trials, statins have been well tolerated, and muscle pain-related side effects were similar to those caused by placebo. An exchange of statins may be beneficial, although all statins have been associated with muscle pain. In some patients, a reduction of dose is worth trying, especially in primary prevention [54]. Statins have been linked also to digestive problems, mental fuzziness, and glucose metabolism, and they may rarely cause liver damage. The influence of the diabetogenic action of statins is still unclear. Despite these observations, the CVD preventive benefit of statin treatment outweighs the CVD risk associated with the development of new diabetes [55]. There is good evidence that statins given late in life to people at risk for vascular disease do not prevent cognitive decline or dementia [56]. Statins can cause transient elevation of liver enzymes, which has led to the unnecessary cessation of these substances prematurely [57]. Coenzyme Q10 (CoQ10) is widely used as a dietary supplement, and one of its roles is to act as an antioxidant. Decreased levels have been shown in diseased myocardium and in Parkinson disease. Farnesyl pyrophosphate is a critical intermediate for CoQ10 synthesis, and blockage of this mechanism may be important in statin myopathy. Supplementation with CoQ10 has been reported to be beneficial in treating hypertension, statin myopathy, heart failure, and problems associated with chemotherapy; however, this use of CoQ10 as a supplement has not been confirmed in randomized controlled clinical trials [58].

In conclusion, recent analyses and randomized controlled trials have been published confirming that the cardiovascular benefits of statin therapy in patients for whom it is recommended by current guidelines greatly outweigh the risks of side effects [59]. The Cholesterol Treatment Trialists Collaboration meta-analysis showed that for each 1 mmol/L reduction in LDL, major vascular events (myocardial infarction, coronary artery disease death, or any stroke or coronary revascularization) were reduced by 22% and total mortality was reduced by 10% over 5 years [59].

Extracting Categories From Raw Data

Our data set contains 1986 unique sentences about cholesterol and statins. Of the 1986 sentences, 1041 (52.42%) were labeled by medical experts as credible, 551 (27.74%) as neutral, and 394 (19.84%) as noncredible. We have reviewed the compliance of the assessments in the noncredible class with the annotation protocol. As a result, of the 394 noncredible annotations, 72 (18.3%) were discarded as noncompliant. The following are some examples of sentences erroneously annotated as noncredible:

Chapter 5. Articles comprising the thesis

“Why are they putting patient lives at risk?” Sentence is a question and should be labeled as neutral.

“Researchers chose 30 studies in total to analyze.” Sentence does not contain any medical terms and should be labeled as neutral.

“They [statins] work by blocking an enzyme called HMG-CoA reductase, which makes your body much slower at synthesizing cholesterol.” Sentence contains factually true statement and should be labeled as credible.

Finally, of the 1986 sentences, we identified 322 (16.21%) as noncredible. We extracted 18 claim categories, which represented 61.5% (198/322) of all noncredible sentences. The process of claim category extraction involved the following steps:

1. The annotator examined all the sentences from the noncredible class one by one.
2. If a sentence matched an already existing category, it was assigned to that category; otherwise, a new category was created.
3. After processing all the sentences, categories with only 1 sentence were merged into a Miscellaneous category that contained the remaining 29.5% (95/322) of the noncredible sentences.

We also compared the compliance of the extracted claim categories with current medical guidelines and knowledge. The category counts are presented in [Table 4](#), and these categories are listed and explained in [Table 5](#)

Table 4. The number of occurrences of a particular claim category within the *cholesterol* and *statins* subset of sentences.

Claim category	Number of occurrences	Is related claim factually incorrect?	Is category based on the content or on the form?
Miscellaneous	95	N/A ^a	Form
(stat) Side effects	43	Yes	Content
(chol) Not an indicator of CVD ^b risk	25	Yes	Content
Diet as good as drugs	22	Yes	Form
(chol) Too low is harmful	18	Yes	Content
Lifestyle changes are enough	15	Yes	Content
Big pharma	14	Yes	Content
Inflammation theory	14	Yes	Content
(stat) Cause diabetes	13	Yes	Content
(stat) Not needed	10	Yes	Content
(chol) Makes cells and protects nerves	8	No	Content
(stat) Not effective	7	Yes	Content
(stat) Prescription based solely on (chol) level	7	Yes	Content
Detailed data	7	N/A	Form
(stat) Cause cognitive impairment	6	Yes	Content
(stat) Not studied enough	6	Yes	Content
High HDL ^c neutralizes high LDL ^d	6	No	Content
Harmful CoQ10 ^e loss	4	Yes	Content
(chol) Consumption not an issue	3	Yes	Content
Lifestyle versus statins	2	Yes	Content
No liver function monitoring	2	Yes	Content

^aN/A: not applicable.

^bCVD: cardiovascular disease.

^cHDL: high-density lipoprotein.

^dLDL: low-density lipoprotein.

^eCoQ10: Coenzyme Q10.

Chapter 5. Articles comprising the thesis

Table 5. Claim category and explanations of claim categories extracted manually from all noncredible sentences from the *cholesterol* and *statins* topic.

Claim category	Claim explanation
(stat) Side effects	Statins' side effects outweigh the benefits
(chol) Not an indicator of CVD ^a risk	Total cholesterol is not an indicator of CVD
Diet as good as drugs	Aggregation of different dietary interventions to lower cholesterol, triglycerides, or sugars
(chol) Too low is harmful	Too low cholesterol level is harmful
Lifestyle changes are enough	People can lower cholesterol level just by developing good habits and eating a proper diet
Big pharma	People (eg, physicians and pharmaceutical company workers) make considerable profit through prescribing statins
Inflammation theory	It is inflammation that causes CVD, not excessive cholesterol level; cholesterol is an effect, not a cause
(stat) Cause diabetes	Statins increase the risk of diabetes
(stat) Not needed	Statins are given to healthy people who do not need them
(chol) Makes cells and protects nerves	Cholesterol produces hormones that make body cells and protect nerves
(stat) Not effective	Statins do not fulfill their role in reducing the risk of CVD
(stat) Prescription based solely on (chol) level	Statin prescription is based solely on total cholesterol level
Detailed data	Sentences contain detailed data, for example, "LDL ^b cholesterol level should not exceed 200 md/dL"
(stat) Cause cognitive impairment	Statin consumption causes different forms of cognitive impairment (including memory loss and slow information processing)
(stat) Not studied enough	Statins' effectiveness is not studied enough
High HDL ^c neutralizes high LDL	HDL is a so-called good cholesterol, whereas LDL is a so-called bad cholesterol; high levels of the former neutralize negative consequences of high levels of the latter
Harmful CoQ10 ^d loss	Statin-related CoQ10 loss is harmful
(chol) Consumption not an issue	People should not worry about cholesterol consumption
Lifestyle versus statins	Lifestyle changes are more effective ways to prevent CVDs than statin consumption
No liver function monitoring	Monitoring of liver function tests is no longer recommended in patients on statin therapy
Miscellaneous	None of the above

^aCVD: cardiovascular disease.

^bLDL: low-density lipoprotein.

^cHDL: high-density lipoprotein.

^dCoQ10: Coenzyme Q10.

Of the 322 noncredible sentences, 198 (61.5%) fall into specific claim categories. Most of the categories have at least 6 examples that spread across different documents. We have designated categories with only 2 or 3 occurrences as separate because the entire noncredible class is relatively small and finding even a few similar sentences may indicate that the claim is being duplicated on the web.

Of the 95 sentences that did not fall into any claim category, we identified 9 (9%) that bear the hallmarks of a conspiracy theory, 7 (7%) containing reasoning based on anecdotal evidence, and 9 (9%) containing misleading statistical reporting:

- Conspiracy theory (referring to groups of interests such as prostatin vs antistatin researchers): "Ironically, prostatin researchers themselves are the ones who are guilty of cherry-picking."
- Anecdotal evidence: "What's worse, my doctor has never asked if I smoke cigarettes, exercise regularly, or eat a healthy diet."

- Misleading statistical evidence: "OK, maybe the benefits of taking a statin are small, but many smart doctors say a reduction of five-tenths or six-tenths of 1% is worthwhile."

As part of qualitative analysis, we compared 2 sets of clusters: automatically created versus manually created. We were able to select sentences that contain similar words and statements but differ in the narrative details that skewed the experts' judgments. We have identified 4 types of false and misleading narratives that occur frequently in the noncredible class. These narratives are as follows:

1. Slippery slope: The sentence is factually true, but the consequences of the presented fact are exaggerated. Example:

Hence, while the drug might synergise with a statin to prevent a non-fatal (or minor) heart attack, it seems to increase the risk of some other equally life-threatening pathology, resulting in death.

Chapter 5. Articles comprising the thesis

Cholesterol also helps in the formation of your memories and is vital for neurological function.

2. Hedging: The sentence is factually incorrect, but there is a part of it that softens the overtone of the presented statement. Example:

However, cholesterol content should be less of a concern than fat content. [CRED]

Coenzyme Q10 supplements may help prevent statin side effects in some people, though more studies are needed to determine any benefits of taking it. [CRED]

The FDA warns on statin labels that some people have developed memory loss or confusion while taking statins. [CRED]

3. Suggested negative consequences: The sentence is mostly factually true, but given the context of the expert's experience, there is a risk that the presented information may lead the patient to act contrary to current medical guidelines. Examples:

For starters, statin drugs deplete your body of coenzyme Q10 (CoQ10), which is beneficial to heart health and muscle function.

Cholesterol is a waxy, fatty steroid that your body needs for things like: cell production.

4. Twisting words: the presence of a single word changes the overtone of the sentence. Examples:

Statins may slightly increase the risk for Type 2 diabetes, a condition that can lead to heart disease or stroke. [CRED]

For example, it may be enough to eat a nutritious diet, exercise regularly, and avoid smoking tobacco products. [NONCRED]

versus

Eating a healthy diet and doing regular exercise can help lower the level of cholesterol in your blood. [CRED]

Discussion

Principal Findings

The results of our experiments show that applying the active annotation paradigm for credibility assessment in the medical domain produces measurable gains in terms of the use of medical experts' time. Active annotation allows us to raise the number of noncredible statements annotated by medical experts by 30% on average, within a fixed time and monetary budget. Annotation of medical information cannot be crowdsourced because it requires the deep and broad domain knowledge of medical experts and their time is expensive. We regard the problem of prohibitively expensive annotation costs as the main obstacle to the broad use of machine learning models in the evaluation of the credibility of web-based medical resources. Our proposal is a step toward a significant lowering of these costs.

However, there is still room for improvement. Our qualitative analysis shows that most of the noncredible sentences can be classified into a limited number of categories. The subset of

approximately 200 noncredible sentences from the *cholesterol and statins* subdomain can be divided into 18 categories, each representing approximately one false statement. These 18 categories fall into 61.5% (198/322) of the total number of all sentences labeled in full accordance with the annotation protocol. This indicates the importance of precise semantic clustering. More accurate clustering helps to detect noncredible sentences faster. It also enables the tagging of clusters with topic-related labels by nonexperts for later reviewing by medical experts and, as a result, the even more useful sentence ranking. In other words, it might be possible to use crowdsourcing to some extent during preprocessing and include an expert in the loop in the main annotation pipeline, further reducing the annotation costs.

Another conclusion that we drew from the qualitative analysis concerns the precision of the semantic similarity measure based on sentence embeddings. The method captures well the overall theme of the sentence but often misses the stance of the presented claim. This error is understandable because the stance in the medical domain is often expressed through subtle sentence modifications, as listed in the *Results* section. Sentence embeddings also struggle with finding a good representation of the form of the sentence—whether it is a supposition, a question, or a statement. Recognition of the form of the sentence can improve the accuracy of classification of neutral sentences that do not require medical expert annotation.

Finally, the qualitative analysis has revealed 4 distinct narratives present in noncredible sentences. Although our analysis was limited to the topic of cholesterol and statins, we feel that these narratives are more general in nature and may apply broadly to false medical information on other topics. If this hypothesis is confirmed, it may be possible to develop machine learning models for these narratives (eg, a model searching for instances of hedging expressions or words capable of twisting the stance of the sentence). Tagging these narratives during credibility annotation may not only increase the precision of sentence classifiers built upon such data sets, but, most importantly, also help disambiguate experts' labeling process.

Conclusions and Future Work

With the web quickly becoming one of the primary sources of the first medical information for the general public [60], the ability to distinguish between credible and noncredible information is indispensable. Financial interests of the alternative medicine community, combined with the rising distrust of the medical establishment, produce voluminous corpora of medical information of questionable quality. Of note, too many people fall prey to medical misinformation because it becomes increasingly harder to tell credible content from harmful deceit.

A possible solution to the problem of medical information source credibility is external certification. In our experiments, we correlated medical experts' labels with HON labels. The certification certainly works because only 18% (240/1333) of the sentences originating from HON-certified websites were classified by our experts as noncredible. However, obtaining the certificate is not simple, the certification process is long, and the entire framework does not scale well. This scalability

Chapter 5. Articles comprising the thesis

problem demonstrates the bottleneck of any approach used for checking the credibility of medical content—the availability and time of medical professionals who need to be involved in the evaluation. In our work, we have taken the approach of optimizing the use of the time spent by experts on credibility evaluation of medical web content. The main goal of our future work will be the improvement and extension of this approach using active annotation and active learning methods.

In contrast, an ambitious goal would be to replace medical experts' evaluations with an automated credibility evaluation system. Such a system would use advanced natural language processing and machine classification algorithms. The results of our research demonstrate the challenges that would need to be overcome to make this possible.

The computational linguistic community is currently divided into 2 opposing camps: those who attribute *understanding of meaning* to language models and those who do not [61]. Despite the recent successes of modern language models such as Generative Pre-trained Transformer 3, the evidence seems to support a more cautious position. Indeed, a language model trained only on the form (raw text) cannot capture the true meaning of the text. The meaning, in this context, should be understood as the relationship between the linguistic form and the communicative intent of the speaker.

Our case goes beyond the learning of the meaning of sentences. As we have shown in this paper, there is an additional layer of complexity introduced by the notion of credibility of a statement to a user. Many machine learning solutions focus on the identification of factual flaws when addressing misinformation. However, fact-checking is not enough in the medical information domain. Often one encounters fake news and disinformation woven around factually true statements. We have seen time and time again medical experts using contextual information when assigning labels denoting sentence credibility. Most often they would take into account the most probable course of action taken by a patient who consumes medical information. Because of this mechanics of annotation, the relationship between sentence credibility and sentence truthfulness becomes

ambiguous, further complicating the shape of the decision boundary between credible and noncredible medical statements.

This observation leads us to an important conclusion about the design of information-processing pipelines for medical content credibility evaluation. The first step is the compilation of large, high-quality data sets for machine learning model training. The active annotation approach presented in this paper allows doubling the number of sentences annotated by medical experts per cost unit (time or monetary). This, in turn, results in larger and more comprehensive training data sets. As a side effect, active annotation produces topical clusters of sentences, which can be used in 2 ways: (1) by allowing nonexpert annotators (whose time is far less expensive) to preprocess large batches of sentences to be reviewed by medical experts and (2) by reducing the cognitive stress of expert annotators due to the removal of context switching.

These 2 effects combined can further enhance the annotation process and increase the volume of annotated data. We also plan to extend the scope of the data set by covering more topics and providing more annotations.

The second step toward the support of medical content credibility evaluation would be the investigation of statistical models' efficacy for automatic classification of medical sentences as either credible or noncredible. Having an accurate classifier of medical sentence credibility, we might develop machine-assisted methods for finding consensus among human annotators, for example, by correlating human annotations with the confidence scores of the classifier. Finally, we would like to pursue active annotation in the light of 2 frameworks. Bayesian reasoning provides a set of tools for modeling individual annotators' beliefs about annotated data. Expectation maximization, in contrast, allows finding the best approximations (or maximum a posteriori estimates) of the unknown point credibility scores from empirical data. We see several possibilities of including the active annotation step in the iterative processes of Bayesian inference or expectation maximization.

Acknowledgments

The research leading to these results has received funding from the European Economic Area Financial Mechanism 2014-2021 (project registration number: 2019/35/J/HS6/03498).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Queries used to retrieve articles.

[\[DOCX File, 13 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

List of article URLs.

[\[DOCX File, 36 KB-Multimedia Appendix 2\]](#)

References

Chapter 5. Articles comprising the thesis

1. Zarocostas J. How to fight an infodemic. *Lancet* 2020 Feb 29;395(10225):676 [FREE Full text] [doi: [10.1016/S0140-6736\(20\)30461-X](https://doi.org/10.1016/S0140-6736(20)30461-X)] [Medline: [32113495](https://pubmed.ncbi.nlm.nih.gov/32113495/)]
2. 5G conspiracy theories prosper during the coronavirus pandemic. *Snopes*. URL: <https://www.snopes.com/news/2020/04/09/5g-conspiracy-theories-prosper-during-the-coronavirus-pandemic> [accessed 2021-11-01]
3. Jablonka A, Happel C, Grote U, Schleenvoigt BT, Hampel A, Dopfer C, et al. Measles, mumps, rubella, and varicella seroprevalence in refugees in Germany in 2015. *Infection* 2016 Dec;44(6):781-787. [doi: [10.1007/s15010-016-0926-7](https://doi.org/10.1007/s15010-016-0926-7)] [Medline: [27449329](https://pubmed.ncbi.nlm.nih.gov/27449329/)]
4. Medical credibility corpus. GitHub. URL: https://github.com/alenabozny/medical_credibility_corpus [accessed 2021-11-17]
5. Sørensen K, Pelikan JM, Röthlin F, Ganahl K, Slonska Z, Doyle G, HLS-EU Consortium. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *Eur J Public Health* 2015 Dec;25(6):1053-1058 [FREE Full text] [doi: [10.1093/eurpub/ckv043](https://doi.org/10.1093/eurpub/ckv043)] [Medline: [25843827](https://pubmed.ncbi.nlm.nih.gov/25843827/)]
6. Keleher H, Hagger V. Health literacy in primary health care. *Aust J Prim Health* 2007 Jul 15;13(2):24-30 [FREE Full text] [doi: [10.1071/PY07020](https://doi.org/10.1071/PY07020)]
7. Cheng C, Dunn M. Health literacy and the internet: a study on the readability of Australian online health information. *Aust N Z J Public Health* 2015 Aug 25;39(4):309-314. [doi: [10.1111/1753-6405.12341](https://doi.org/10.1111/1753-6405.12341)] [Medline: [25716142](https://pubmed.ncbi.nlm.nih.gov/25716142/)]
8. Trethewey SP. Strategies to combat medical misinformation on social media. *Postgrad Med J* 2020 Jan 15;96(1131):4-6 [FREE Full text] [doi: [10.1136/postgradmedj-2019-137201](https://doi.org/10.1136/postgradmedj-2019-137201)] [Medline: [31732511](https://pubmed.ncbi.nlm.nih.gov/31732511/)]
9. Ecker UK, O'Reilly Z, Reid JS, Chang EP. The effectiveness of short-format refutational fact-checks. *Br J Psychol* 2020 Feb 02;111(1):36-54 [FREE Full text] [doi: [10.1111/bjop.12383](https://doi.org/10.1111/bjop.12383)] [Medline: [30825195](https://pubmed.ncbi.nlm.nih.gov/30825195/)]
10. Nyhan B, Reifler J. When corrections fail: the persistence of political misperceptions. *Polit Behav* 2010 Mar 30;32(2):303-330. [doi: [10.1007/s11109-010-9112-2](https://doi.org/10.1007/s11109-010-9112-2)]
11. Horne BD, Gruppi M, Adali S. Trustworthy misinformation mitigation with soft information nudging. In: Proceedings of the 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA). 2019 Presented at: 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA); Dec 12-14, 2019; Los Angeles, CA, USA URL: <https://ieeexplore.ieee.org/document/9014346> [doi: [10.1109/tps-isa48467.2019.00039](https://doi.org/10.1109/tps-isa48467.2019.00039)]
12. Dhoju S, Rony MM, Kabir MA, Hassan N. Differences in health news from reliable and unreliable media. In: Proceedings of the WWW '19: The Web Conference. 2019 Presented at: WWW '19: The Web Conference; May 13-17, 2019; San Francisco USA. [doi: [10.1145/3308560.3316741](https://doi.org/10.1145/3308560.3316741)]
13. Fernández-Pichel M, Losada DE, Pichel JC, Elsweiler D. Reliability prediction for health-related content: a replicability study. In: *Advances in Information Retrieval*. Cham: Springer; 2021.
14. Afsana F, Kabir MA, Hassan N, Paul M. Automatically assessing quality of online health articles. *IEEE J Biomed Health Inform* 2021 Feb;25(2):591-601. [doi: [10.1109/jbhi.2020.3032479](https://doi.org/10.1109/jbhi.2020.3032479)] [Medline: [33079686](https://pubmed.ncbi.nlm.nih.gov/33079686/)]
15. Augenstein I, Lioma C, Wang D, Lima LC, Hansen C, Hansen C, et al. MultiFC: a real-world multi-domain dataset for evidence-based fact checking of claims. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov 2019; Hong Kong, China URL: <https://aclanthology.org/D19-1475> [doi: [10.18653/v1/d19-1475](https://doi.org/10.18653/v1/d19-1475)]
16. Wang WY. "Liar, Liar Pants on Fire": a new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017 Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers); Jul 2017; Vancouver, Canada URL: <https://aclanthology.org/P17-2067> [doi: [10.18653/v1/p17-2067](https://doi.org/10.18653/v1/p17-2067)]
17. Rashkin H, Choi E, Jang JY, Volkova S, Choi Y. Truth of varying shades: analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; Sep, 2017; Copenhagen, Denmark. [doi: [10.18653/v1/d17-1317](https://doi.org/10.18653/v1/d17-1317)]
18. Tchechmedjiev A, Fafalios P, Boland K, Gasquet M, Zloch M, Zapilko B, et al. ClaimsKG: a knowledge graph of fact-checked claims. In: *The Semantic Web – ISWC 2019*. Cham: Springer; Oct 2019.
19. Hassan N, Arslan F, Li C, Tremayne M. Toward automated fact-checking: detecting check-worthy factual claims by ClaimBuster. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017 Presented at: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13 - 17, 2017; Halifax NS Canada. [doi: [10.1145/3097983.3098131](https://doi.org/10.1145/3097983.3098131)]
20. Karmakharm T, Aletras N, Bontcheva K. Journalist-in-the-loop: continuous learning as a service for rumour analysis. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations; Nov 2019; Hong Kong, China. [doi: [10.18653/v1/d19-3020](https://doi.org/10.18653/v1/d19-3020)]

Chapter 5. Articles comprising the thesis

21. Samuel H, Zaïane O. MedFact: towards improving veracity of medical information in social media using applied machine learning. In: *Advances in Artificial Intelligence*. Cham: Springer International Publishing; 2018.
22. Yilmaz Z, Yang W, Zhang H, Lin J. Cross-domain modeling of sentence-level evidence for document retrieval. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019 Presented at: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Nov, 2019; Hong Kong, China URL: <https://aclanthology.org/D19-1352> [doi: [10.18653/v1/d19-1352](https://doi.org/10.18653/v1/d19-1352)]
23. Chen D, Zhang S, Zhang X, Yang K. Cross-lingual passage re-ranking with alignment augmented multilingual BERT. *IEEE Access* 2020 Dec 1;8:213232-213243. [doi: [10.1109/access.2020.3041605](https://doi.org/10.1109/access.2020.3041605)]
24. Ghenai A, Mejova Y. Catching Zika fever: application of crowdsourcing and machine learning for tracking health misinformation on Twitter. In: *Proceedings of the 2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 2017 Presented at: *IEEE International Conference on Healthcare Informatics (ICHI)*; Aug 23-26, 2017; Park City, UT, USA. [doi: [10.1109/ichi.2017.58](https://doi.org/10.1109/ichi.2017.58)]
25. Shepperd S, Charnock D. Why DISCERN? *Health Expect* 1998 Nov 04;1(2):134-135 [FREE Full text] [doi: [10.1046/j.1369-6513.1998.0112a.x](https://doi.org/10.1046/j.1369-6513.1998.0112a.x)] [Medline: [11281867](https://pubmed.ncbi.nlm.nih.gov/11281867/)]
26. Moults B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. *Health Expect* 2004 Jun;7(2):165-175 [FREE Full text] [doi: [10.1111/j.1369-7625.2004.00273.x](https://doi.org/10.1111/j.1369-7625.2004.00273.x)] [Medline: [15117391](https://pubmed.ncbi.nlm.nih.gov/15117391/)]
27. Bunge M, Mühlhauser I, Steckelberg A. What constitutes evidence-based patient information? Overview of discussed criteria. *Patient Educ Couns* 2010 Mar;78(3):316-328. [doi: [10.1016/j.pec.2009.10.029](https://doi.org/10.1016/j.pec.2009.10.029)] [Medline: [20005067](https://pubmed.ncbi.nlm.nih.gov/20005067/)]
28. Working Group GPGI. Good practice guidelines for health information. *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen* 2016;110-111:e1-e8 [FREE Full text] [doi: [10.1016/j.zefq.2016.01.004](https://doi.org/10.1016/j.zefq.2016.01.004)]
29. Keselman A, Smith CA, Murcko AC, Kaufman DR. Evaluating the quality of health information in a changing digital ecosystem. *J Med Internet Res* 2019 Feb 08;21(2):e11129 [FREE Full text] [doi: [10.2196/11129](https://doi.org/10.2196/11129)] [Medline: [30735144](https://pubmed.ncbi.nlm.nih.gov/30735144/)]
30. Marinelli F, Cervone A, Tortoreto G, Stepanov EA, Di Fabrizio G, Riccardi G. Active annotation: bootstrapping annotation lexicon and guidelines for supervised NLU learning. *Proc Interspeech* 2019:574-578. [doi: [10.21437/interspeech.2019-2537](https://doi.org/10.21437/interspeech.2019-2537)]
31. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems*. 2013 Presented at: *Proceedings of the 26th International Conference on Neural Information Processing Systems*; Dec 5 - 10, 2013; Lake Tahoe Nevada. [doi: [10.5555/2999792.2999959](https://doi.org/10.5555/2999792.2999959)]
32. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014 Presented at: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Oct, 2014; Doha, Qatar URL: <https://aclanthology.org/D14-1162> [doi: [10.3115/v1/d14-1162](https://doi.org/10.3115/v1/d14-1162)]
33. Levy O, Goldberg Y. Dependency-based word embeddings. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2014 Presented at: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*; Jun, 2014; Baltimore, Maryland URL: <https://aclanthology.org/P14-2050> [doi: [10.3115/v1/p14-2050](https://doi.org/10.3115/v1/p14-2050)]
34. Nickel M, Kiela D. Poincaré embeddings for learning hierarchical representations. *arXiv*. 2017. URL: <https://arxiv.org/abs/1705.08039> [accessed 2021-11-17]
35. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *ArXiv.org*. 2017 Dec. URL: <https://arxiv.org/abs/1607.04606> [accessed 2021-11-17]
36. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997 Nov 15;9(8):1735-1780. [doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735)] [Medline: [9377276](https://pubmed.ncbi.nlm.nih.gov/9377276/)]
37. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014 Presented at: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Oct, 2014; Doha, Qatar URL: <https://aclanthology.org/D14-1179> [doi: [10.3115/v1/d14-1179](https://doi.org/10.3115/v1/d14-1179)]
38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *arXiv*. 2017. URL: <https://arxiv.org/abs/1706.03762> [accessed 2021-11-17]
39. Cer D, Yang Y, Kong S, Hua N, Limtiaco N, St. John R, et al. Universal sentence encoder. *arXiv*. 2018. URL: <https://arxiv.org/abs/1803.11175> [accessed 2021-11-17]
40. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. *ArXiv.org*. 2020. URL: <https://arxiv.org/abs/2005.14165> [accessed 2021-11-17]
41. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. *ArXiv.org*. 2018. URL: <https://arxiv.org/abs/1810.04805> [accessed 2021-11-17]
42. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. *ArXiv.org*. 2019. URL: <https://arxiv.org/abs/1908.10084> [accessed 2021-11-17]

Chapter 5. Articles comprising the thesis

43. Chicco D. Siamese neural networks: an overview. *Methods Mol Biol* 2021;2190:73-94. [doi: [10.1007/978-1-0716-0826-5_3](https://doi.org/10.1007/978-1-0716-0826-5_3)] [Medline: [32804361](https://pubmed.ncbi.nlm.nih.gov/32804361/)]
44. Sentence (linguistics). Wikipedia. URL: [https://en.wikipedia.org/wiki/Sentence_\(linguistics\)](https://en.wikipedia.org/wiki/Sentence_(linguistics)) [accessed 2021-11-17]
45. Lorenz-Spreen P, Mønsted BM, Hövel P, Lehmann S. Accelerating dynamics of collective attention. *Nat Commun* 2019 Apr 15;10(1):1759 [FREE Full text] [doi: [10.1038/s41467-019-09311-w](https://doi.org/10.1038/s41467-019-09311-w)] [Medline: [30988286](https://pubmed.ncbi.nlm.nih.gov/30988286/)]
46. Nabožny A, Balcerzak B, Koržinek D. Enriching the context: methods of improving the non-contextual assessment of sentence credibility. In: *Web Information Systems Engineering – WISE 2019*. Cham: Springer; 2019.
47. Krishna K, Murty MN. Genetic K-means algorithm. *IEEE Trans Syst Man Cybern B Cybern* 1999;29(3):433-439. [doi: [10.1109/3477.764879](https://doi.org/10.1109/3477.764879)] [Medline: [18252317](https://pubmed.ncbi.nlm.nih.gov/18252317/)]
48. Yuan C, Yang H. Research on K-Value selection method of K-means clustering algorithm. *J* 2019 Jun 18;2(2):226-235. [doi: [10.3390/j2020016](https://doi.org/10.3390/j2020016)]
49. Liu Y, Myle O, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. ArXiv.org. 2019. URL: <https://arxiv.org/abs/1907.11692> [accessed 2021-11-17]
50. Raunak V. Simple and effective dimensionality reduction for word embeddings. ArXiv.org. 2017. URL: <https://tinyurl.com/vf29aah8> [accessed 2021-11-17]
51. Mu J, Bhat S, Viswanath P. All-but-the-top: simple and effective postprocessing for word representations. ArXiv.org. 2017. URL: <https://arxiv.org/abs/1702.01417> [accessed 2021-11-17]
52. Alhmod EN, Barazi R, Fahmi A, Abdu A, Higazy A, Elhajj M. Critical appraisal of the clinical practice guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular Riskuropean Society of Cardiology (ESC) and European Atherosclerosis Society (ESC/EAS) 2019 guidelines. *J Pharm Health Serv Res* 2020 Nov;11(4):423-427. [doi: [10.1111/jphs.12371](https://doi.org/10.1111/jphs.12371)]
53. Ferri N, Corsini A. Clinical pharmacology of statins: an update. *Curr Atheroscler Rep* 2020 Jun 03;22(7):26. [doi: [10.1007/s11883-020-00844-w](https://doi.org/10.1007/s11883-020-00844-w)] [Medline: [32494971](https://pubmed.ncbi.nlm.nih.gov/32494971/)]
54. Pergolizzi Jr JV, Coluzzi F, Colucci RD, Olsson H, LeQuang JA, Al-Saadi J, et al. Statins and muscle pain. *Expert Rev Clin Pharmacol* 2020 Mar 27;13(3):299-310. [doi: [10.1080/17512433.2020.1734451](https://doi.org/10.1080/17512433.2020.1734451)] [Medline: [32089020](https://pubmed.ncbi.nlm.nih.gov/32089020/)]
55. Yandrapalli S, Malik A, Guber K, Rochlani Y, Pemmasani G, Jasti M, et al. Statins and the potential for higher diabetes mellitus risk. *Expert Rev Clin Pharmacol* 2019 Sep 31;12(9):825-830. [doi: [10.1080/17512433.2019.1659133](https://doi.org/10.1080/17512433.2019.1659133)] [Medline: [31474169](https://pubmed.ncbi.nlm.nih.gov/31474169/)]
56. McGuinness B, Craig D, Bullock R, Passmore P. Statins for the prevention of dementia. *Cochrane Database Syst Rev* 2016 Jan 04(1):CD003160. [doi: [10.1002/14651858.CD003160.pub3](https://doi.org/10.1002/14651858.CD003160.pub3)] [Medline: [26727124](https://pubmed.ncbi.nlm.nih.gov/26727124/)]
57. Shrestha A, Mulmi A, Munankarmi R. Statins and abnormal liver enzymes. *S D Med* 2019 Jan;72(1):12-14. [Medline: [30849222](https://pubmed.ncbi.nlm.nih.gov/30849222/)]
58. Saha SP, Whyne TF. Coenzyme Q-10 in human health: supporting evidence? *South Med J* 2016 Jan;109(1):17-21. [doi: [10.14423/SMJ.0000000000000393](https://doi.org/10.14423/SMJ.0000000000000393)] [Medline: [26741866](https://pubmed.ncbi.nlm.nih.gov/26741866/)]
59. Cholesterol Treatment Trialists' (CTT) Collaboration, Baigent C, Blackwell L, Emberson J, Holland LE, Reith C, et al. Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. *Lancet* 2010 Nov 13;376(9753):1670-1681 [FREE Full text] [doi: [10.1016/S0140-6736\(10\)61350-5](https://doi.org/10.1016/S0140-6736(10)61350-5)] [Medline: [21067804](https://pubmed.ncbi.nlm.nih.gov/21067804/)]
60. Sun Y, Zhang Y, Gwizdzka J, Trace CB. Consumer evaluation of the quality of online health information: systematic literature review of relevant criteria and indicators. *J Med Internet Res* 2019 May 02;21(5):e12522 [FREE Full text] [doi: [10.2196/12522](https://doi.org/10.2196/12522)] [Medline: [31045507](https://pubmed.ncbi.nlm.nih.gov/31045507/)]
61. Bender EM, Koller A. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020 Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jul, 2020; Online p. 5185-5198. [doi: [10.18653/v1/2020.acl-main.463](https://doi.org/10.18653/v1/2020.acl-main.463)]

Abbreviations

- BERT:** Bidirectional Encoder Representations from Transformers
 - CoQ10:** Coenzyme Q10
 - CVD:** cardiovascular disease
 - HON:** Health on the Net
 - LDL:** low-density lipoprotein
 - sBERT:** sentence-Bidirectional Encoder Representations from Transformers
-
-

Chapter 5. Articles comprising the thesis

Edited by C Lovis; submitted 26.11.20; peer-reviewed by W Xu, M Jordan-Marsh, A Hidki, S Nagavally; comments to author 09.02.21; revised version received 29.03.21; accepted 24.09.21; published 26.11.21

Please cite as:

Nabożny A, Balcerzak B, Wierzbicki A, Morzy M, Chlabicz M

Active Annotation in Evaluating the Credibility of Web-Based Medical Information: Guidelines for Creating Training Data Sets for Machine Learning

JMIR Med Inform 2021;9(11):e26065

URL: <https://medinform.jmir.org/2021/11/e26065>

doi: [10.2196/26065](https://doi.org/10.2196/26065)

PMID:

©Aleksandra Nabożny, Bartłomiej Balcerzak, Adam Wierzbicki, Mikołaj Morzy, Małgorzata Chlabicz. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 26.11.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

- 5.3 Article 3 "Focus On Misinformation: Improving Medical Experts' Efficiency Of Misinformation Detection" (WISE 2021, 140 pts., Awarded as 'Best Student Paper runner-up')**

Focus On Misinformation: Improving Medical Experts' Efficiency Of Misinformation Detection

Aleksandra Nabożny¹[0000-0001-9534-142X], Bartłomiej
Balcerzak²[0000-0003-0881-5362], Mikołaj Morzy³[0000-0002-2905-9538], and
Adam Wierzbicki²[0000-0003-0075-7030]

¹ Gdańsk University of Technology

² Polish-Japanese Institute of Information Technology

³ Poznań University of Technology

Abstract. Fighting medical disinformation in the era of the global pandemic is an increasingly important problem. As of today, automatic systems for assessing the credibility of medical information do not offer sufficient precision to be used without human supervision, and the involvement of medical expert annotators is required. Thus, our work aims to optimize the utilization of medical experts' time. We use the dataset of sentences taken from online lay medical articles. We propose a general framework for filtering medical statements that do not need to be manually verified by medical experts. The results show the gain in fact-checking performance of expert annotators on capturing misinformation by the factor of 2.2 on average. In other words, our framework allows medical experts to fact-check and identify over two times more non-credible medical statements in a given time interval without applying any changes to the annotation flow.

Keywords: e-health · misinformation · text-mining · human-in-the-loop · credibility assessment · natural language processing · machine learning

1 Introduction

The spread of medical misinformation on the World Wide Web has become a significant social problem. We face a global "infodemic" of dubious medical claims, distrust in medical science, conspiracy theories, and outright medical falsehoods circulating in social media. The recent SARS-CoV-2 pandemic has exacerbated the existing problem of low confidence in medical institutions, pharmaceutical companies, and governmental agencies responsible for public health [13, 18]. At the same time, we observe a growing trend of relying on online health information for self-treatment [5]. Given the possible consequences of using online health advice ungrounded in medical science, the task of assessing the credibility of online health information becomes pressing.

Distinguishing between reliable and unreliable online health information poses a substantial challenge for lay Internet users [1]. Labeling source websites as either credible or non-credible is not sufficient as false claims can be a part of an

article originating from a credible source and vice versa. Often, credible medical statements can serve as the camouflage for disinformation woven into otherwise factually correct statements. Even subtle changes to the overtone, wording, or strength of a medical statement can change its meaning, for instance, by exaggerating the side effects of a drug or by conflating relative and absolute risks of a medical procedure. As an example, consider the following phrase: *"Aspirin should not be consumed during pregnancy"*. This phrase is generally true but does not apply when an early pregnancy is at the risk of miscarriage when consuming small doses of aspirin can significantly lower the risk.

Even experienced medical professionals find it challenging to assess the truthfulness of online medical information. What is considered to be "true" in the domain of medicine is often subject to a very complex context. This context is provided by external medical knowledge and clinical practice. Medical professionals often focus on the possible impact of health information on the choices made by patients rather than evaluate the factual correctness of a statement. In other words, a factually correct statement may still inflict health damage on patients when presented mischievously or in isolation. The phrase *"For starters, statin drugs deplete your body of coenzyme Q10 (CoQ10), which is beneficial to heart health and muscle function"*, despite factual correctness, would raise objections from medical professionals as it may discourage a patient from taking statins. In this example, the expert uses external knowledge from their clinical practice that for patients requiring statin therapy, its benefits far outweigh the potential risks associated with coenzyme Q10 deficiency. This additional context of online health information evaluation makes it extremely difficult to frame the task in terms of machine learning.

An additional problem that arises when evaluating online health information stems from the involvement of human judges. Whether these are annotators who curate training data for statistical models or subject matter experts (SMEs) who provide final scores, it is paramount that trained medical practitioners perform these tasks. Unfortunately, data labeling for online health information assessment, to the large extent, cannot be crowd-sourced due to the unique competencies required to provide the ground truth labels. Over-worked medical practitioners struggle to secure the time required for debunking online medical falsehoods and cannot keep up with the flood of online medical misinformation.

Scarce human resources are the bottleneck stifling the development of automatic online health information assessment methods. To address this issue, we propose to frame the problem of online health information evaluation as a machine learning problem, with the business objective being the optimization of the utilization of medical experts' time. Firstly, though, we have to change the definition of the machine learning task. As we have stated above, assessing the truthfulness of medical statements is subjective, context-dependent, and challenging. Instead, we propose to develop machine learning models that assess the *credibility* of medical statements. We define a medical statement to be credible if the statement is in accord with current medical knowledge and does not entice a patient to make harmful health-related decisions or to inspire actions

contrary to the current medical guidelines. We do not try to discover the intention of an author of online health information. Thus we use the general term "misinformation" to represent both malicious and unintentional deception.

The business objective of optimizing the utilization of medical experts' time has yet to be framed in terms of an objective function driving the training of statistical models. We treat the time budget allocated by a medical expert to debunking online medical information as a fixed value. Similarly, we treat the average time required by a medical expert to evaluate a single medical statement as a fixed value. The only intervention that can influence the utilization of medical experts' time is the re-ranking of medical statements for annotation. We propose to focus medical experts' attention on statements that are possibly non-credible and contain medical misinformation. This, in turn, requires the development of methods for the automatic discovery of credible statements. The objective function is to maximize the recall of credible medical statements at a fixed high precision threshold. In this way, we can extract a large set of medical statements which are guaranteed to contain credible medical information due to fixed precision, and remove these statements from the queue of statements for human annotation, allowing medical experts to focus their limited time on the discovery of non-credible statements. Our experiments show that this approach increases the utilization of medical experts' time by the factor of 2.

Our main contributions presented in this paper include:

- introduction of the general approach for the optimization of the utilization of human annotators' time using machine learning,
- evaluation of the approach using the task of annotating online medical information credibility,
- developing a set of statistical classifiers for assessing the credibility of medical statements with the precision ranging from 83.5% to 98.6% for credible statements across ten different medical topics,
- developing a new method of data and label augmentation for improving the accuracy of the credibility classifiers,
- experimental evaluation of the augmentation method proving its efficacy.

2 Related Work

There are multiple strategies for improving the credibility of online health information. They include information corrections, both automatically-generated and user-generated [4], and the manipulation of the visual appeal and presentation of medical information [8]. A recent meta-analysis [23] shows, however, that the average effect of correction of online health information on social media is of weak to moderate magnitude. The authors point out that interventions are more effective in cases when misinformation distributed by news organizations is debunked by medical experts. When misinformation is circulated on social media by peers, or when non-experts provide corrections, interventions have low impact.

The approaches to automatic classification of online medical misinformation differ depending on the media and content type. Most studies employ content analysis, social network analysis, or experiments, drawing from disciplinary paradigms [24]. Online medical misinformation can be effectively classified by using so-called peripheral-level features [29] which include linguistic features (length of a post, presence of a picture, inclusion of an URL, content similarity with the main discussion thread), sentiment features (both corpus-based and language model-based), and behavioral features (discussion initiation, interaction engagement, influential scope). Peripheral-level features proved to be useful for detecting the spread of false medical information during the Zika virus epidemic [7, 22]. Stylistic features can be used to identify hoaxes presented as genuine news articles and promoted on social media [19]. Along with identifying hoaxes, it is possible to identify social media users who are prone to disseminating these hoaxes among peers [9]. An applied machine learning-based approach, called *MedFact*, is proposed in [21], where the authors present an algorithm for trusted medical information recommendation. The *MedFact* algorithm relies on keyword extraction techniques to assess the factual accuracy of statements posted in online health-related forums.

More advanced methods of online medical information evaluation include video analysis (extracting medical knowledge from YouTube videos [15]), detecting misinformation based on multi-modal features (both text and graphics [25]), and website topic classification. The latter approach was successfully applied by [2, 14] using topic analysis (either Latent Dirichlet Annotation or Term-Frequency). In addition, Afsana *et al.* use linguistic features, such as word counts, named entities, semantic coherence of articles, the Linguistic Inquiry Word Count (LIWC), and external metrics such as citation counts and Web ranking of a document.

We consider the full article’s credibility prediction as burdened with source bias, as well as not precise enough to perform the targeted decision explanations. That is why, instead of the articles, we chose to classify smaller chunks of text (triplets of sentences, in particular). In previous approaches, the classifiers rated entire documents. For example, in the study evaluating entire articles [2], they were assessed against 10 criteria, none of which directly determines whether the content is credible or not. Our method differs from the approaches presented in the literature earlier in two important aspects: we leverage the context of medical expert’s annotation by data and label augmentation, and we modify the objective function to optimize for the recall of the positive class given the fixed precision threshold.

3 Methods

3.1 Dataset

Our dataset consists of over 10 000 sentences extracted from 247 online medical articles. The articles have been manually collected from health-related websites.

Table 1. Number of sentences from each class by the topic.

Category	Topic	CRED	NEU	NONCRED
Cardiology	Antioxidants	375	175	144
Cardiology	Heart supplements	221	124	78
Cardiology	Cholesterol and statins	1058	565	406
Gynecology	Cesarean section vs. natural birth	275	53	31
Pediatrics	Children & antibiotics	298	52	82
Pediatrics	Diet and Autism	236	71	124
Pediatrics	Steroids for kids	560	101	40
Pediatrics	Vaccination	730	223	309
Pediatrics	Allergy testing	790	398	214
Psychiatry	Psychiatry	1194	676	402

The choice of major categories (cardiology, gynecology, psychiatry, and pediatrics) has been dictated by the availability of medical experts participating in the experiment. After consulting with medical experts, we have selected certain topics known to produce controversy in online social networks. For each topic, we have collected a diversified sample of articles presenting contradicting views (either supportive or contrarian) and we have extracted statements for manual evaluation by medical experts. The dataset is open-sourced and publicly available ⁴. For the detailed description of the dataset, we refer the reader to [16].

Nine medical experts took part in the experiment, including 2 cardiologists, 1 gynecologist, 3 psychiatrists, and 3 pediatricians. All experts have completed 6-years medical studies and then a 5-year residency program. The experts were paid for a full day of work (approximately 8 hours each). Each medical expert had at least 10 years of clinical experience, except for the gynecologist who was a resident doctor. We have accepted his participation in the experiment due to his status as a Ph.D. candidate in the field of medicine. One of the psychiatrists held a Ph.D. in medical sciences. Given the high qualifications of participants, we consider their judgments as the ground truth for medical statement evaluation. The experts were allowed to browse certified medical information databases throughout the experiment. Each expert evaluated the credibility of medical statements only within their specialization.

Collected online articles were automatically divided into sentences and presented to the medical experts in random order. Sentence segmentation has been done using the dependency parser from the `spaCy` text processing library. Since input text follows closely the general-purpose news style, the default `spaCy` processing pipeline produces very robust sentence segmentation. Along with each sentence we have displayed a limited number of automatically extracted keywords. If the medical expert decided that a sentence could not have been assessed due to insufficient context, he or she could have expanded the annotation view by showing preceding and succeeding sentences. Each medical expert was asked to annotate approximately 1000 sentences. Medical experts evaluated the

⁴ https://github.com/alenabozny/medical_credibility_corpus

credibility of sentences with the following set of labels and the corresponding instructions:

- **CRED** (credible) — a sentence is reliable, does not raise major objections, contains verifiable information from the medical domain.
- **NONCRED** (non-credible) — a sentence contains false or unverifiable information, contains persuasion contrary to current medical recommendations, contains outdated information.
- **NEU** (neutral) — a sentence does not contain factual information (e.g., is a question) or is not related to medicine.

Table 1 presents the number of sentences in each class summarized by category and topic. Within the four larger topical categories (cardiology, gynecology, psychiatry, or pediatrics), our dataset is divided into smaller subsets (topics). Considering these topics separately dramatically improves the performance of the classifiers. However, some topics included in the dataset were too small for training a classifier. Thus, we do not consider them further in this article.

3.2 Data augmentation

The annotation of the dataset by medical experts has revealed the importance of context for providing a label (see Table 2). Over 25% of non-credible sentences required the surrounding context of one sentence, with 20% of credible sentences and 12% neutral sentences requiring similar context. To provide this context for statistical models, we have decided to transform single sentences into sequences of consecutive non-overlapping triplets of sentences. Since individual sentences have already been labeled by medical experts, we have transferred ground truth sentence labels to triplet labels in the following way:

- **negative**: a triplet is negative if any of the sentences constituting the triplet has the label **NONCRED**,
- **positive**: a triplet is positive if all of the sentences constituting the triplet are either **CRED** or **NEU**.

Figure 1 depicts the idea of label transfer applied to the dataset.

Table 2. Number m of surrounding sentences needed to understand the context and evaluate the credibility of a sentence for credible, non-credible, neutral, and all sentences.

m	credible [%]	non-credible [%]	neutral [%]	all [%]
0	80.07	71.27	88.30	80.43
1	18.83	26.60	11.03	18.39
> 1	0.18	0.37	0.04	0.18

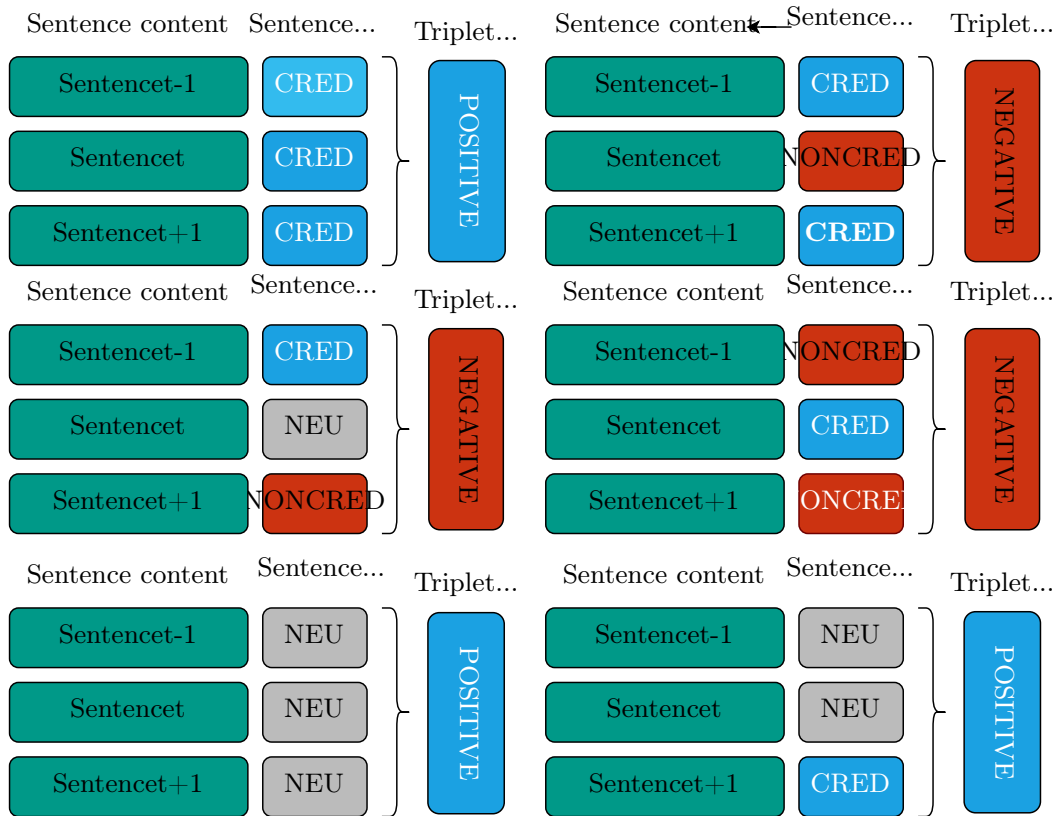


Fig. 1. Label transfer for augmented data

Example of a positive triplet (from "Statins & cholesterol"):

"Not smoking could add nearly 10 years and quitting increases life expectancy by reducing the chances of emphysema, many cancers, and heart disease. Although my doctor checks my cholesterol every year, it remains low and taking a statin will have a very small, if any, effect on my life expectancy. What's worse, my doctor has never asked if I smoke cigarettes, exercise regularly, or eat a healthy diet."

Example of a negative triplet (from "Statins & cholesterol"):

"OK, maybe the benefits of taking a statin are small, but many smart doctors say a reduction of five-tenths or six-tenths of 1% is worthwhile. Yet the few published observations on people over the age of 70 do not show any statistically significant statin-related reductions in deaths from any cause. Of course, not everyone is like me."

3.3 Feature set

Features that have been selected for credibility classification purposes are based on the qualitative analysis of the dataset concerning the findings reported in Section 2. The ultimate number of features varies between categories. The feature set has been created manually and feature selection methods have been used to remove non-informative features. The choice of traditional NLP features has been deliberate as we want to maintain the explainability of credibility classification models. Also, somewhat contrary to popular belief, initial experiments have shown that these traditional features are superior to sentence embeddings computed using BERT [6]. It remains to be seen if using a language model fine-tuned to the medical domain [3] would make the embeddings a better source of features for the credibility classifier.

Uncased TF-IDF (number of features: varying from 920 to 4103) Bag of words, n-gram, term frequency (TF), term frequency inverted document frequency (TF-IDF) are the most commonly used textual features in natural language processing [28]. In this work, we chose TF-IDF values to account for the importance of each word. We use the Python package `spaCy` to perform sentence tokenization.

Dependency tree-labels count (number of feaures: up to 45) Overly complex sentences have a higher probability of containing the hedging part than simple sentences (the base of a sentence may contain a factually false statement, but the other part would soften its overtone so that it seems credible). Thus, we count the base elements of dependency trees to model the potential existence of such phenomena.

Named Entities counter (number of features: up to 18) There are some indicators of conspiratorial and/or science-skeptical language (hence the popularity of using agent-action-target triples in the study of conspiratorial narratives [20]) Those narratives may be captured by counting named entities of specified categories, such as false authority (PERSON), Big Pharma blaming (ORGANIZATION, PRODUCT), distrust to renowned institutions (ORGANIZATION), facts and statistics (NUMBER). In the experiment we have used the NER labeling scheme available in the English language model offered by the `spaCy` library.

Polarity and subjectivity (number of features: 2) Sentiment analysis is a broadly-used feature set for misinformation detection classifiers. It has been used, for example, for detecting anti- and pro-vaccine news headlines [27]. Highly polarized and/or emotional language can indicate misinformation.

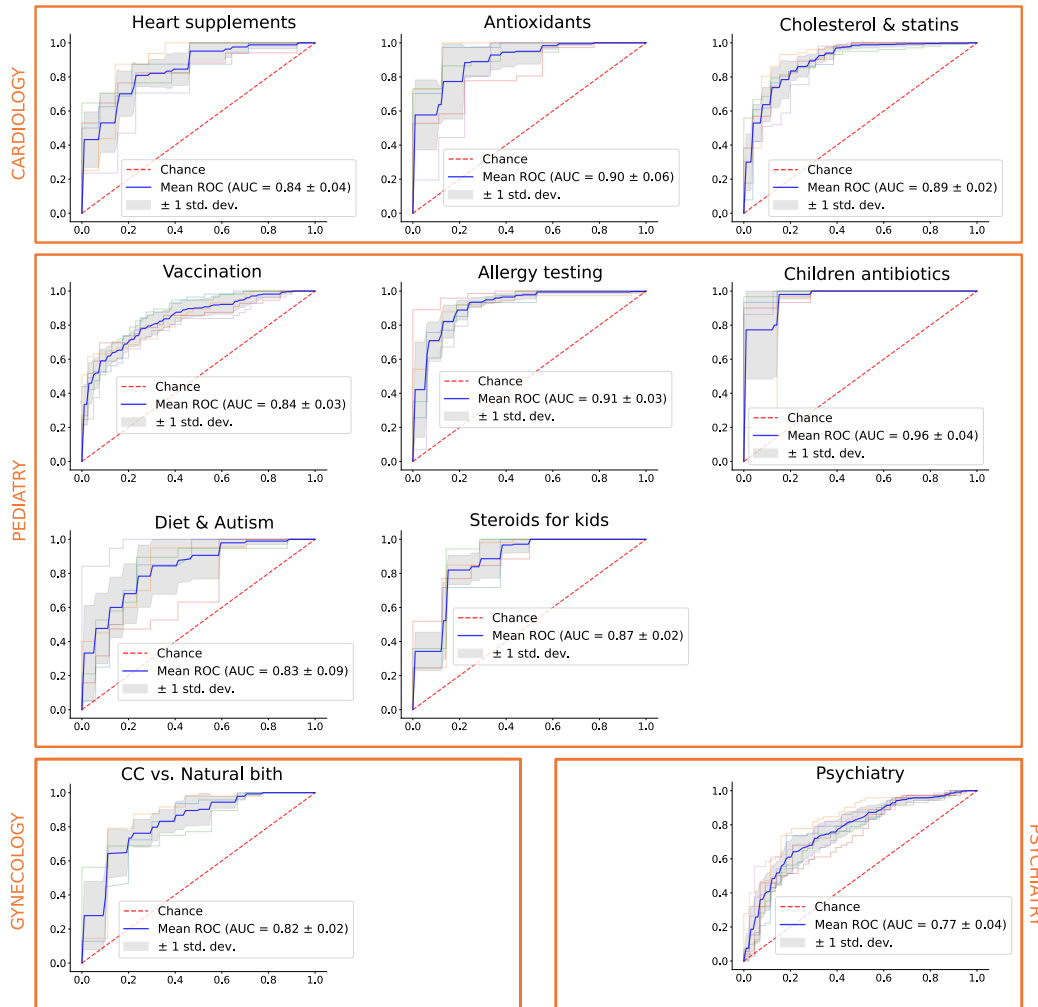


Fig. 2. ROC curves of cross-validated classification results for each medical topic.

LIWC (number of features: 93) Aggressive, overly optimistic, advertising language (e.g. for a drug or novel therapy) or other patterns can affect the credibility of textual information [12]. The LIWC offers a corpus-based sentiment analysis approach by counting words in different emotion categories. Empirical results using LIWC demonstrate its ability to detect meaning in emotionality. In addition, it has been employed to extract the sentiment features for the detection of misinformation in online medical videos [11]. LIWC provides features regarding emotional dimensions, the formality of the language, spatial and temporal features, as well as structural information (e.g. word per sentence count).

3.4 Feature selection and model training

The workflow for training statistical models is identical for each topic and includes two steps: feature selection and model selection. Feature selection is per-

formed using Logistic Regression and Recursive Feature Elimination (RFE) [10]. RFE conducts a backward selection of features, starting from a predictive model using all available features. For each feature, the importance score is computed, and the least important feature is removed. The model is retrained with remaining features and the procedure is repeated until the desired number of features remains. We use Logistic Regression as the baseline model for RFE, limiting the number of features to 30% of the number of samples in a given topic. Due to the lack of space we cannot provide a detailed description of selected feature sets for each medical category, but we will include this information in the extended version of the paper. In this paper we also assume that the list of topics is known in advance and that each sentence is already assigned to a topic. This, of course, raises the question of practical applicability of our method when the topic of an article is unknown. Recent advances in automatic medical subdomain classification [26] suggest that the topic of the article can be successfully extracted from the text.

For training the model we use the TPOT library [17]. TPOT uses a genetic algorithm to optimize the workflow consisting of feature pre-processing, model selection, and parameter optimization, by evolving a population of workflows and implementing mutation and cross-over operators for workflows. To constrain the space of considered models we use Logistic Regression, XGBoost, and the Multi-layer Perceptron as the initial pool of available models. The optimization is driven by the F_1 measure.

4 Results

The main objective of our method is the maximization of the utilization of medical experts' time when annotating online medical statements. We optimize statistical models to find credible statements, thus increasing the number of non-credible statements that can be presented to medical experts. The results below analyze the efficiency of trained statistical models in finding credible statements. Recall from Section 3.2 that statistical models are trained on a binary dataset consisting of positive (credible and neutral) and negative (non-credible) triplets of sentences.

Figure 2 presents ROC curves for cross-validation. The number of folds depends on the number of samples in a given topic. Based on the ROC curves we have empirically adjusted the cutoff threshold for each classifier's prediction of the positive class. Our goal was to maximize the recall of the positive class while preserving fixed high precision for the positive class. In other words, samples that fall above the cutoff threshold are assumed to contain only credible or neutral sentences, and will not be presented to medical experts for manual evaluation. We have selected the cutoff threshold for each topic using the following criteria:

- the difference between the proportion of true negative samples and the proportion of negative samples in the entire test set should be maximized, with minimum variance,

Chapter 5. Articles comprising the thesis

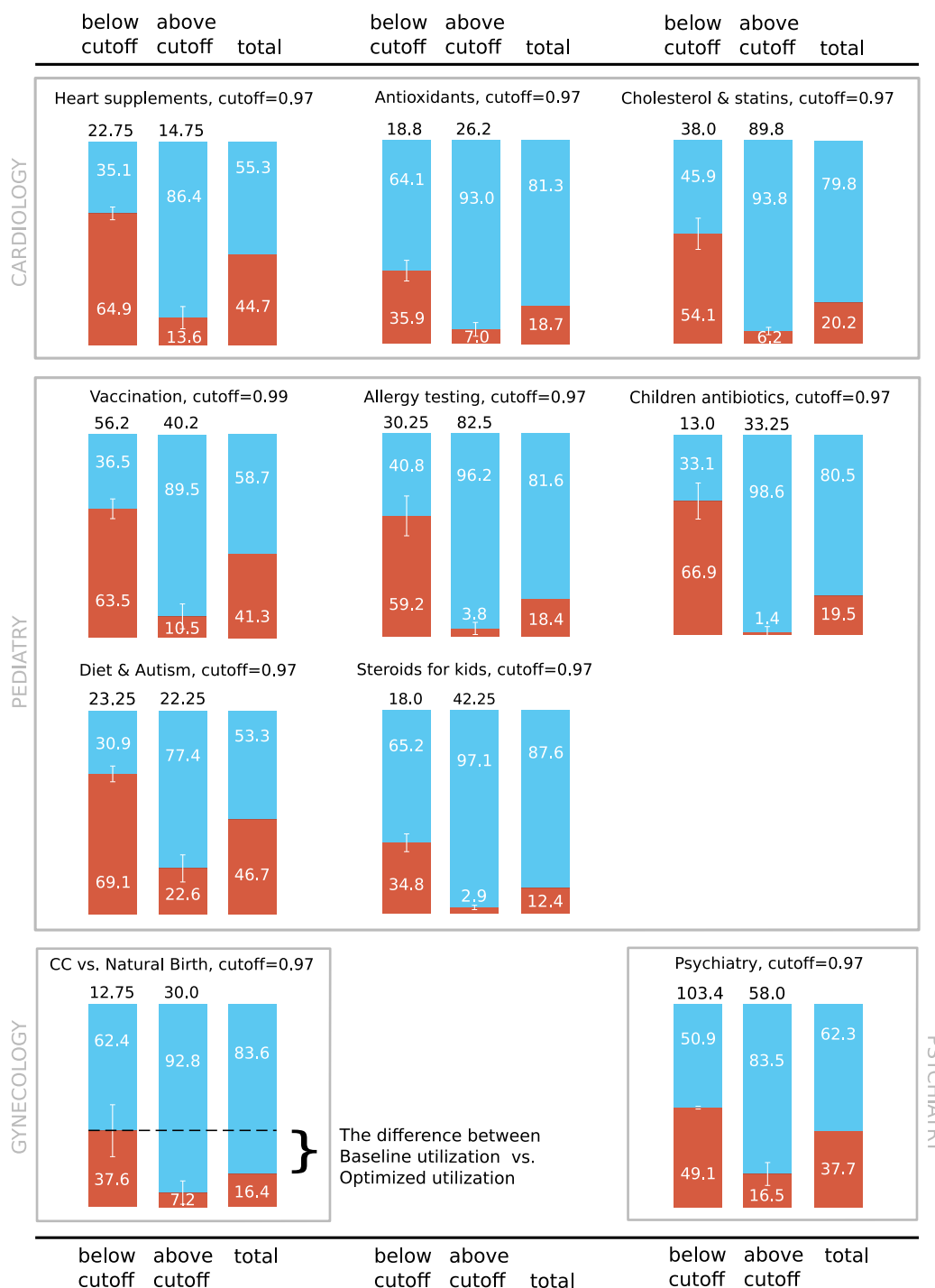


Fig. 3. Cross-validated proportions of positive and negative samples a) below the cutoff b) above the cutoff c) in the entire test set. Black labels indicate the mean number of samples in each group

Chapter 5. Articles comprising the thesis

12 A. Nabożny *et al.*

- the precision for the true positive class should be maximized,
- the number of samples above the cutoff should be maximized.

The results of the cutoff filtering are presented in Figure 3. For each topic, we show the distribution of positive and negative samples in the entire topic (the *total* column) and in the subsets defined by the cutoff. For instance, there are 44.7% of negative samples and 55.3% of positive samples in the *Heart supplements* topic. The subset of samples defined by the cutoff point of 0.97 contains only 13.6% of negative samples, and the remaining subset contains 64.9% of negative samples. In other words, by removing the samples above the cutoff threshold from manual experts’ evaluation we are increasing the number of negative samples that the experts may annotate from 44.7% to 64.9%. We refer to the proportion of negative samples in the topic as the *baseline utilization*, and the proportion of negative samples after the intervention (i.e., below the cutoff threshold) as the *optimized utilization*.

Table 3 presents the main results of our experiment. We report baseline and optimized utilization, the difference in percentage points, and the factor of improvement of medical experts’ time utilization.

Table 3. Comparison of baseline and optimized utilization of medical experts’ time.

Category	Baseline utilization [%]	Optimized utilization [%]	pp. diff	factor
Heart supplements	44.7	64.9	↑ 20.2	1.5
Antioxidants	18.7	35.9	↑ 17.2	1.9
Cholesterol & statins	20.2	54.1	↑ 33.9	2.7
Vaccination	41.3	63.5	↑ 22.2	1.5
Allergy testing	18.4	59.2	↑ 40.8	3.2
Children antibiotics	19.5	66.9	↑ 47.4	3.4
Diet & Autism	46.7	69.1	↑ 22.4	1.5
Steroids for kids	12.4	34.8	↑ 22.4	2.8
CC vs. Natural Birth	16.4	37.6	↑ 21.2	2.3
Psychiatry	37.7	49.1	↑ 11.4	1.3
mean	-	-	↑ 25.9	2.2

5 Discussion

Evaluation of the credibility of online medical information is a very challenging task due to the subjective assessment of credibility, and the specialized medical knowledge required to perform the evaluation [16]. Fully automatic classification of online medical information as credible or non-credible is not a viable solution due to the complex externalities involved in such classification. For the foreseeable future, keeping a human judge in the annotation loop is a necessity. At

the same time, qualified human judges are the scarcest resource and their time must be utilized efficiently. Previous approaches to automatically assessing the credibility of medical texts did not take into account the need to weave a human judge into the real-time verification process.

In our work, we present a framework for the optimization of the utilization of medical experts' time when evaluating the credibility of online medical information. To prioritize the evaluation of non-credible information by medical experts, we train classifiers that can filter out credible and neutral medical claims with very high precision exceeding 90% for most medical topics considered in our study (vaccination, allergy testing, children antibiotics, steroids for kids, antioxidants, cholesterol & statins, and C-section vs. natural birth).

Table 3 depicts the key benefit for the potential human-in-the-loop fact-checking system that our solution provides — an increase in the probability that a medical expert will encounter a non-credible medical statement in the annotation batch. As we can see, for all topics the improvement in the utilization of medical experts' time is substantial. The average improvement over all topics is 25.9 percentage points, which means that within the same amount of time and at the same average time needed to annotate a single sentence, medical experts using our method annotate over two times as many non-credible medical statements on average. It is a "pure win" since this improvement does not require any changes to either the annotation protocol or the annotation interface, we simply make much better use of the valuable experts' time allocated to data annotation.

6 Conclusions and Future Work

One limitation of our method is a small number of statements that contain misinformation, but would not be seen by experts. However, we need to keep in mind that medical experts may not be able to see all statements anyway, as their time and attention are limited and may not be enough to process all suspicious information.

In a realistic use-case, medical experts would continually evaluate a stream of statements derived from the ever-growing set of online articles on medical and health topics, as well as information from social media. Our method allows increasing the efficiency of misinformation detection by debunking medical experts, who will discover more than twice as much misinformation without increasing the time spent on evaluation (or the number of evaluating experts), and without introducing any changes to the annotation workflow. Our method can be regarded as a universal filter for medical web content.

In our future work, we will focus on gathering more data by introducing the demo expert crowd-sourcing system in a few medical universities. We will put special emphasis on the iterative process of adjusting proper annotation protocol and professional training for medical students to gain the annotation accuracy as close as possible to the experts (medical practitioners with at least a few years of experience), thus further reducing costs of expert medical credibility annotation.

References

1. Abramczuk, K., Kąkol, M., Wierzbicki, A.: How to Support the Lay Users Evaluations of Medical Information on the Web? (2016). https://doi.org/10.1007/978-3-319-40349-6_1
2. Afsana, F., Kabir, M.A., Hassan, N., Paul, M.: Automatically Assessing Quality of Online Health Articles. *IEEE Journal of Biomedical and Health Informatics* **25**(2) (2 2021). <https://doi.org/10.1109/JBHI.2020.3032479>
3. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323* (2019)
4. Bode, L., Vraga, E.K.: See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication* **33**(9), 1131–1140 (9 2018). <https://doi.org/10.1080/10410236.2017.1331312>
5. Chen, Y.Y., Li, C.M., Liang, J.C., Tsai, C.C.: Health information obtained from the internet and changes in medical decision making: questionnaire development and cross-sectional survey. *Journal of medical Internet research* **20**(2), e47 (2018)
6. 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)
7. Dito, F.M., Alqadhi, H.A., Alasaadi, A.: Detecting Medical Rumors on Twitter Using Machine Learning. In: 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies, 3ICT 2020. Institute of Electrical and Electronics Engineers Inc. (12 2020). <https://doi.org/10.1109/3ICT51146.2020.9311957>
8. Ebnali, M., Kian, C.: Nudge Users to Healthier Decisions: A Design Approach to Encounter Misinformation in Health Forums (2020). https://doi.org/10.1007/978-3-030-20500-3_1
9. Ghenai, A., Mejova, Y.: Fake Cures. *Proceedings of the ACM on Human-Computer Interaction* **2**(CSCW) (11 2018). <https://doi.org/10.1145/3274327>
10. Guyon, I., Weston, J., Barnhill, S.: Gene Selection for Cancer Classification using Support Vector Machines. *Tech. rep.* (2002)
11. Hou, R., Perez-Rosas, V., Loeb, S., Mihalcea, R.: Towards Automatic Detection of Misinformation in Online Medical Videos. In: 2019 International Conference on Multimodal Interaction. ACM, New York, NY, USA (10 2019). <https://doi.org/10.1145/3340555.3353763>
12. Jensen, M.L., Averbek, J.M., Zhang, Z., Wright, K.B.: Credibility of Anonymous Online Product Reviews: A Language Expectancy Perspective. *Journal of Management Information Systems* **30**(1) (7 2013). <https://doi.org/10.2753/MIS0742-1222300109>
13. Latkin, C.A., Dayton, L., Yi, G., Konstantopoulos, A., Boodram, B.: Trust in a COVID-19 vaccine in the U.S.: A social-ecological perspective. *Social Science & Medicine* **270** (2 2021). <https://doi.org/10.1016/j.socscimed.2021.113684>
14. Li, J.: Detecting False Information in Medical and Healthcare Domains: A Text Mining Approach (2019). https://doi.org/10.1007/978-3-030-34482-5_21
15. Liu, X., Zhang, B., Susarla, A., Padman, R.: YouTube for Patient Education: A Deep Learning Approach for Understanding Medical Knowledge from User-Generated Videos. *ArXiv Computer Science* (7 2018)
16. Nabożny, A., Balcerzak, B., Wierzbicki, A., Morzy, M.: Digging for the truth: the case for active annotation in evaluating the credibility of online medical information. *JMIR Preprints* (11 2020)

17. Olson, R.S., Urbanowicz, R.J., Andrews, P.C., Lavender, N.A., Kidd, L.C., Moore, J.H.: Automating Biomedical Data Science Through Tree-Based Pipeline Optimization (2016). https://doi.org/10.1007/978-3-319-31204-0_9, <https://epistasislab.github.io/tpot/citing/>
18. Pollard, M.S., Lois M. Davis: Decline in Trust in the Centers for Disease Control and Prevention During the COVID-19 Pandemic. Tech. rep. (2021). <https://doi.org/https://doi.org/10.7249/RRA308-12>
19. Purnomo, M.H., Sumpeno, S., Setiawan, E.I., Purwitasari, D.: Biomedical Engineering Research in the Social Network Analysis Era: Stance Classification for Analysis of Hoax Medical News in Social Media. *Procedia Computer Science* **116** (2017). <https://doi.org/10.1016/j.procs.2017.10.049>
20. Samory, M., Mitra, T.: 'The Government Spies Using Our Webcams': The Language of Conspiracy Theories in Online Discussions. *Proceedings of the ACM on Human-Computer Interaction* **2**(CSCW) (11 2018). <https://doi.org/10.1145/3274421>
21. Samuel, H., Zaïane, O.: MedFact: Towards Improving Veracity of Medical Information in Social Media Using Applied Machine Learning (2018). https://doi.org/10.1007/978-3-319-89656-4_9
22. Sicilia, R., Lo Giudice, S., Pei, Y., Pechenizkiy, M., Soda, P.: Twitter rumour detection in the health domain. *Expert Systems with Applications* **110** (11 2018). <https://doi.org/10.1016/j.eswa.2018.05.019>
23. Walter, N., Brooks, J.J., Saucier, C.J., Suresh, S.: Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. *Health Communication* (8 2020). <https://doi.org/10.1080/10410236.2020.1794553>
24. Wang, Y., McKee, M., Torbica, A., Stuckler, D.: Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine* **240** (11 2019). <https://doi.org/10.1016/j.socscimed.2019.112552>
25. Wang, Z., Yin, Z., Argyris, Y.A.: Detecting medical misinformation on social media using multimodal deep learning (12 2020)
26. Weng, W.H., Waghlikar, K.B., McCray, A.T., Szolovits, P., Chueh, H.C.: Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC medical informatics and decision making* **17**(1), 1–13 (2017)
27. Xu, Z., Guo, H.: Using Text Mining to Compare Online Pro- and Anti-Vaccine Headlines: Word Usage, Sentiments, and Online Popularity. *Communication Studies* **69**(1), 103–122 (1 2018). <https://doi.org/10.1080/10510974.2017.1414068>
28. Zhang, X., Ghorbani, A.A.: An overview of online fake news: Characterization, detection, and discussion. *Information Processing and Management* **57**(2) (3 2020). <https://doi.org/10.1016/j.ipm.2019.03.004>
29. 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) (1 2021). <https://doi.org/10.1016/j.ipm.2020.102390>

- 5.4 Article 4 "Improving medical experts' efficiency of misinformation detection: an exploratory study" (World Wide Web, 100 pts.)



Improving medical experts' efficiency of misinformation detection: an exploratory study

Aleksandra Nabożny¹ · Bartłomiej Balcerzak² · Mikołaj Morzy^{2,3} · Adam Wierzbicki² · Pavel Savov² · Kamil Warpechowski²

Received: 19 January 2022 / Revised: 3 May 2022 / Accepted: 4 July 2022
© The Author(s) 2022

Abstract

Fighting medical disinformation in the era of the pandemic is an increasingly important problem. Today, automatic systems for assessing the credibility of medical information do not offer sufficient precision, so human supervision and the involvement of medical expert annotators are required. Our work aims to optimize the utilization of medical experts' time. We also equip them with tools for semi-automatic initial verification of the credibility of the annotated content. We introduce a general framework for filtering medical statements that do not require manual evaluation by medical experts, thus focusing annotation efforts on non-credible medical statements. Our framework is based on the construction of filtering classifiers adapted to narrow thematic categories. This allows medical experts to fact-check and identify over two times more non-credible medical statements in a given time interval without applying any changes to the annotation flow. We verify our results across a broad spectrum of medical topic areas. We perform quantitative, as well as exploratory analysis on our output data. We also point out how those filtering classifiers can be modified to provide experts with different types of feedback without any loss of performance.

Keywords e-health · Misinformation · Text-mining · Human-in-the-loop · Credibility assessment · Natural language processing · Machine learning

This article belongs to the Topical Collection: *Special Issue on Web Information Systems Engineering 2021*

Guest Editors: Hua Wang, Wenjie Zhang, Lei Zou, and Zakaria Maamar

✉ Mikołaj Morzy
Mikolaj.Morzy@put.poznan.pl

¹ Gdańsk University of Technology, Gdańsk, Poland

² Polish-Japanese Academy of Information Technology, Warsaw, Poland

³ Poznań University of Technology, Poznań, Poland

1 Introduction

The spread of medical misinformation on the World Wide Web is a critical problem in today's society. We face a global "infodemic" of outright health-related falsehoods, conspiracy theories, and dubious medical advice circulating in social media. The recent SARS-CoV-2 pandemic has exacerbated the existing distrust in pharmaceutical companies, low confidence in medical science, medical institutions, and governmental agencies responsible for public health [19, 32]. On the other hand, more and more people rely on online health information for self-treatment [6] while lacking the necessary skill to evaluate the credibility of such information. Given the possible consequences of using online health advice ungrounded in medical science, the task of aiding Web users in assessing the credibility of online health information becomes a high priority.

Distinguishing between credible and non-credible online medical information poses a substantial challenge even for experienced medical professionals, and even more so for ordinary Web users whose evaluation may be impacted by cognitive biases or psychological factors [1, 34]. Labeling source websites as either credible or non-credible is insufficient since false claims can be a part of an article originating from a credible source and vice versa. Often, disinformation is woven into factually correct medical statements that serve as camouflage. Even subtle changes to the wording, strength, or overtone of a medical statement can change its meaning, for instance, by exaggerating the side effects of a drug or by conflating relative and absolute risks of a medical procedure. As an example, consider the following phrase: "*Aspirin should not be consumed during pregnancy*". This phrase is generally true but does not apply to an early pregnancy at risk of miscarriage — then, consuming small doses of aspirin can significantly lower the risk. The credibility of medical statements may also significantly depend on the context. For example, the phrase "*For starters, statin drugs deplete your body of coenzyme Q10 (CoQ10), which is beneficial to heart health and muscle function*", despite factual correctness, would raise objections from medical professionals as it may discourage a patient from taking statins. In this example, the expert uses external knowledge from their clinical practice that benefits provided by statins far outweigh the potential risks associated with coenzyme Q10 deficiency for patients requiring statin therapy. This additional context of online health information evaluation makes it extremely difficult to frame the task in terms of machine learning.

Because assessing the truthfulness of medical statements is subjective, context-dependent, and challenging, in our research we formulate a different task for machine learning models: that of credibility evaluation. *Credibility* is a concept that can depend on the truthfulness of information, but also on other aspects, such as the potential for causing harm or misleading persuasion [45]. Consequently, credibility also applies to statements that cannot be directly verified but may still be harmful or misleading.

We define a medical statement to be *non-credible* if the statement is not in accord with current medical knowledge or entices a patient to make harmful health-related decisions, or inspires actions contrary to the current medical guidelines. We also use the general term *misinformation* to represent information that is not credible (regardless of the intention of the author, which may be malicious or benign).

Because of the critical costs of errors, it is paramount that credibility evaluation of health-related Web content is performed or supervised by trained medical practitioners. Those can be annotators who curate training data for statistical models or experts who provide final scores. Unfortunately, such experts' availability, time and attention are scarce resources. Over-worked medical practitioners struggle to secure the time required for

debunking online medical falsehoods and cannot keep up with the flood of online medical misinformation. Scarce human resources, stifling automatic online assessment methods, are the bottleneck. To address this issue, we propose to frame the problem of online health information evaluation as a machine learning problem. We formulate the business objective as the optimization of the utilization of medical experts' time.

Such business objective has yet to be formulated as an objective function driving the training of statistical models. We treat the total time budget of a medical expert for debunking online medical information as a fixed value. Similarly, we treat the average time required by a medical expert to evaluate a single medical statement as a fixed value (the results of our experiments indicate that the average time to evaluate a statement by an expert is about 30 seconds). On average, a medical expert will evaluate a fixed number of statements. Optimizing the expert's time utilization means increasing the proportion of non-credible statements discovered within her/his time budget.

We propose to focus medical experts' attention on statements that are presumably non-credible and contain medical misinformation. This, in turn, requires the development of methods for the automatic discovery of credible statements. The objective is to maximize the precision with respect to non-credible medical statements (precision for the negative class) at a fixed, high precision threshold of filtering credible statements (precision for the positive class). In this way, we can extract a large set of medical statements which are guaranteed to contain credible medical information due to fixed precision and remove these statements from the queue of statements for human annotation, allowing medical experts to focus their limited time on the discovery of non-credible statements. Our experiments show that this approach increases the utilization of medical experts' time by the factor of 2.

Our main contributions presented in this paper include:

- introduction of a general framework to optimize the utilization of medical experts' time when annotating data for downstream training of machine learning models,
- evaluation of the framework on the task of medical misinformation annotation,
- developing a set of filtering classifiers for assessing the credibility of medical statements with the precision ranging from 83.5% to 98.6% for credible statements across ten different medical topics,
- analysis of most significant features that are used by filtering classifiers,
- providing human-interpretable explanations of filtering classifiers.

2 Related work

There are multiple strategies for improving the credibility of online health information. They include information corrections, both automatically-generated and user-generated [4], and the manipulation of the visual appeal and presentation of medical information [11]. A recent meta-analysis [41] shows, however, that the average effect of correction of online health information on social media is of weak to moderate magnitude. The authors point out that interventions are more effective in cases when misinformation distributed by news organizations is debunked by medical experts. When misinformation is circulated on social media by peers, or when non-experts provide corrections, interventions have low impact.

The approaches to automatic classification of online medical misinformation differ depending on the media and content type. Most studies employ content analysis, social network analysis, or experiments, drawing from disciplinary paradigms [42]. Online medical

misinformation can be effectively classified by using so-called peripheral-level features [48] which include linguistic features (length of a post, presence of a picture, inclusion of an URL, content similarity with the main discussion thread), sentiment features (both corpus-based and language model-based), and behavioral features (discussion initiation, interaction engagement, influential scope). Peripheral-level features proved to be useful for detecting the spread of false medical information during the Zika virus epidemic [10, 38]. Stylistic features can be used to identify hoaxes presented as genuine news articles and promoted on social media [33]. Along with identifying hoaxes, it is possible to identify social media users who are prone to disseminating these hoaxes among peers [13]. An applied machine learning-based approach, called *MedFact*, is proposed in [37], where the authors present an algorithm for trusted medical information recommendation. The *MedFact* algorithm relies on keyword extraction techniques to assess the factual accuracy of statements posted in online health-related forums.

More advanced methods of online medical information evaluation include video analysis (extracting medical knowledge from YouTube videos [22]), detecting misinformation based on multi-modal features (both text and graphics [43]), and website topic classification. The latter approach was successfully applied by [2, 21] using topic analysis (either Latent Dirichlet Annotation or Term-Frequency). Alternatively, text summarization may be used for this purpose [3]. In addition, Afsana et al. use linguistic features, such as word counts, named entities, semantic coherence of articles, the Linguistic Inquiry Word Count (LIWC), and external metrics such as citation counts and Web ranking of a document. A similar multi-modal approach is presented by Dhoju et al. [9] to distinguish with very high precision between reliable and unreliable media outlets publishing health-related information. Also, Wagle et al. use multi-modal analysis to evaluate the credibility of health & beauty blogs by analyzing the credibility of the platform, author, and images embedded in the blog [40].

An important aspect of our approach is the interpretability and explainability of filtering classifiers [27]. The description of recent advances in the field of machine learning interpretability is beyond the scope of this paper, interested reader is referred to a very thorough survey of explainable methods for supervised learning [5] and to an excellent book by Molnar [25]. In our work we utilize the Local Interpretable Model-agnostic Explanations (LIME) [35] technique to gain insights into features used by filtering classifiers to identify credible statements. LIME is an example of the black-box approach to model interpretability. Other popular black-box approaches include using Shapley values [24], partial dependence plots [12], and Morris sensitivity analysis [16, 26]. Alternatively, glass-box models can be used to explain algorithmic decisions of machine learning models. The most popular approaches include decision tree-based explainers [15], using Boolean rules to identify target classes [7], and Explainable Boosting Machines [23]. Implementations of many rule-based glass-box models are readily available in the `imodels` library [39].

This paper is the extension of work originally presented during the 22th International Conference on Web Information Systems Engineering WISE'2021 [28]. The original paper focused on improving the utilization of human annotators' time when manually annotating the credibility of medical statements. This work extends previous report in a number of dimensions. We broaden the related literature review, in particular discussing relevant work on explainable machine learning models. We make a detailed report on annotation times recorded during the experiments. We add transformer-based models to the evaluation (BioBERT) and we include the results of these models in the summary of experiments. We present a new section pertaining to the generalization capabilities of tested models. The entire new section is devoted to the issue of explainability of models: we apply

LIME to our filtering classifiers and we compare these explanations with more traditional approach based on Logistic Regression coefficient analysis. Detailed reports on the experimental results (TPOT configurations, Logistic Regression per topic) are included in two appendices.

3 Methods

In this section we introduce the dataset compiled as the result of our project. We describe the annotation protocol and the annotation procedure, albeit in an abridged manner. For the detailed description of the dataset and the annotation process we refer the reader to [29]. We also present the augmentations applied to the data and the set of features used to train filtering classifiers. We conclude the section with the short overview of the training procedure and the introduction of explainable models used in the experiments.

3.1 Dataset

We consider the credibility prediction of the full article as an insufficiently defined task burdened with source bias. That is why, instead of articles, we chose to classify smaller chunks of text (triplets of sentences, in particular). In previous approaches, the classifiers rated entire documents. For example, in the study evaluating entire articles [2], they were assessed against 10 criteria, none of which directly determines whether the content is credible or not. Our method differs from the approaches presented in the literature earlier in two important aspects: we leverage the context of medical expert’s annotation by data and label augmentation, and we modify the objective function to optimize for the recall of the positive class given the fixed precision threshold.

Our dataset consists of over 10000 sentences extracted from 247 online medical articles. The articles have been manually collected from health-related websites. The choice of major categories (cardiology, gynecology, psychiatry, and pediatrics) has been dictated by the availability of medical experts participating in the experiment. After consulting with medical experts, we have selected certain topics known to produce controversy in online social networks. For each topic, we have collected a diversified sample of articles presenting contradicting views (either supportive or contrarian) and we have extracted statements for manual evaluation by medical experts. The dataset is open-sourced and publicly available.¹

Nine medical experts took part in the experiment, including 2 cardiologists, 1 gynecologist, 3 psychiatrists, and 3 pediatricians. All experts have completed 6-years medical studies and then a 5-year residency program. The experts were paid for a full day of work (approximately 8 hours each). Each medical expert had at least 10 years of clinical experience, except for the gynecologist who was a resident doctor. We have accepted his participation in the experiment due to his status as a Ph.D. candidate in the field of medicine. One of the psychiatrists held a Ph.D. in medical sciences. Given the high qualifications of participants, we consider their judgments as the ground truth for medical statement evaluation. The experts were allowed to browse certified medical information databases throughout the experiment. Each expert evaluated the credibility of medical statements only within their specialization.

¹ https://github.com/alenaabozny/medical_credibility_corpus

Chapter 5. Articles comprising the thesis

Collected online articles were automatically divided into sentences and presented to the medical experts in random order. Sentence segmentation has been done using the dependency parser from the `spaCy` text processing library. Since input text follows closely the general-purpose news style, the default `spaCy` processing pipeline produces very robust sentence segmentation. Along with each sentence we have displayed a limited number of automatically extracted keywords. If the medical expert decided that a sentence could not have been assessed due to insufficient context, he or she could have expanded the annotation view by showing preceding and succeeding sentences. Each medical expert was asked to annotate approximately 1000 sentences. Medical experts evaluated the credibility of sentences with the following set of labels and the corresponding instructions:

- CRED (credible) — a sentence is reliable, does not raise major objections, contains verifiable information from the medical domain.
- NONCRED (non-credible) — a sentence contains false or unverifiable information, contains persuasion contrary to current medical recommendations, contains outdated information.
- NEU (neutral) — a sentence does not contain factual information (e.g., is a question) or is not related to medicine.

Table 1 presents the number of sentences in each class summarized by category and topic. Within the four larger topical categories (cardiology, gynecology, psychiatry, or pediatrics), our dataset is divided into smaller subsets (topics). Considering these topics separately dramatically improves the performance of the classifiers. However, some topics included in the dataset were too small for training a classifier. Thus, we do not consider them further in this article.

3.2 Data augmentation

The annotation of the dataset by medical experts has revealed the importance of context for providing a label (see Table 2). Over 25% of non-credible sentences required the surrounding context of one sentence, with 20% of credible sentences and 12% neutral sentences requiring similar context. To provide this context for statistical models, we have decided to transform single sentences into sequences of consecutive non-overlapping triplets of

Table 1 Number of sentences from each class by the topic

Category	Topic	CRED	NEU	NONCRED
Cardiology	Antioxidants	375	175	144
Cardiology	Heart supplements	221	124	78
Cardiology	Cholesterol and statins	1058	565	406
Gynecology	Cesarean section vs. natural birth	275	53	31
Pediatrics	Children & antibiotics	298	52	82
Pediatrics	Diet and Autism	236	71	124
Pediatrics	Steroids for kids	560	101	40
Pediatrics	Vaccination	730	223	309
Pediatrics	Allergy testing	790	398	214
Psychiatry	Psychiatry	1194	676	402

Chapter 5. Articles comprising the thesis

Table 2 Number m of surrounding sentences needed to understand the context and evaluate the credibility of a sentence for credible, non-credible, neutral, and all sentences

m	Credible [%]	Non-credible [%]	Neutral [%]	All [%]
0	80.07	71.27	88.30	80.43
1	18.83	26.60	11.03	18.39
> 1	0.18	0.37	0.04	0.18

sentences. Since individual sentences have already been labeled by medical experts, we have transferred ground truth sentence labels to triplet labels in the following way:

- **negative**: a triplet is negative if any of the sentences constituting the triplet has the label NONCRED,
- **positive**: a triplet is positive if all of the sentences constituting the triplet are either CRED or NEU.

Example of a positive triplet (from "Statins & cholesterol"):

"Not smoking could add nearly 10 years and quitting increases life expectancy by reducing the chances of emphysema, many cancers, and heart disease. Although my doctor checks my cholesterol every year, it remains low and taking a statin will have a very small, if any, effect on my life expectancy. What's worse, my doctor has never asked if I smoke cigarettes, exercise regularly, or eat a healthy diet."

Example of a negative triplet (from "Statins & cholesterol"):

"OK, maybe the benefits of taking a statin are small, but many smart doctors say a reduction of five-tenths or six-tenths of 1% is worthwhile. Yet the few published observations on people over the age of 70 do not show any statistically significant statin-related reductions in deaths from any cause. Of course, not everyone is like me."

3.3 Feature set

Features that have been selected for credibility classification purposes are based on the qualitative analysis of the dataset concerning the findings reported in Section 2. The ultimate number of features varies between categories. The feature set has been created manually and feature selection methods have been used to remove non-informative features. The choice of traditional NLP features has been deliberate as we want to maintain the explainability of filtering classifiers. However, we compare them to the compressed lexical features obtained by the state-of-the-art deep learning language model BioBERT [20] trained on clinical data.

3.3.1 Uncased TF-IDF (number of features: varying from 920 to 4103)

Bag of words, n-gram, term frequency (TF), term frequency inverted document frequency (TF-IDF) are the most commonly used textual features in natural language processing [47].

Chapter 5. Articles comprising the thesis

In this work, we chose TF-IDF values to account for the importance of each word. We use the Python package `spacy` to perform sentence tokenization and lemmatization.

3.3.2 BioBERT vectors (number of features: up to 768)

BioBERT is a pre-trained language representation model for the medical domain. It was designed for linguistic tasks of Medical Entity Recognition, relation extraction, and question answering [8, 49]. The model we use was trained on a combination of general purpose and medical corpora (English Wikipedia, Books Corpus, PubMed Abstracts and PMC full articles). In our work, we decided to use the sentence vectorization module of BioBERT. This module transforms each paragraph in the corpus into a numerical vector. This vector is an aggregation of word embeddings generated for each word in the paragraph by the BioBERT model.

3.3.3 Dependency tree-labels count (number of features: up to 45)

Overly complex sentences have a higher probability to contain the hedging part than simple sentences (the base of a sentence may contain a factually false statement, but the other part would soften its overtone so that it seems credible). Thus, we count the base elements of dependency trees to model the potential existence of such phenomena.

3.3.4 Named entities counter (number of features: up to 18)

There are some indicators of conspiratorial and/or science-skeptical language (hence the popularity of using agent-action-target triples in the study of conspiratorial narratives [36]). Those narratives may be captured by counting named entities of specified categories, such as false authority (PERSON), Big Pharma blaming (ORGANIZATION, PRODUCT), distrust to renowned institutions (ORGANIZATION), facts and statistics (NUMBER). In the experiment we have used the NER labeling scheme available in the English language model offered by the `spacy` library.

3.3.5 Polarity and subjectivity (number of features: 2)

Sentiment analysis is a broadly-used feature set for misinformation detection classifiers. It has been used, for example, for detecting anti- and pro-vaccine news headlines [46]. Highly polarized and/or emotional language can indicate misinformation Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and 13.

3.3.6 LIWC (number of features: 93)

Aggressive, overly optimistic, advertising language (e.g. for a drug or novel therapy) or other patterns can affect the credibility of textual information [18]. The LIWC offers a corpus-based sentiment analysis approach by counting words in different emotion categories. Empirical results using LIWC demonstrate its ability to detect meaning in emotionality. In addition, it has been employed to extract the sentiment features for the detection of misinformation in online medical videos [17]. LIWC provides features regarding emotional dimensions, the formality of the language, spatial and temporal features, as well as structural information (e.g. word per sentence count).

Chapter 5. Articles comprising the thesis

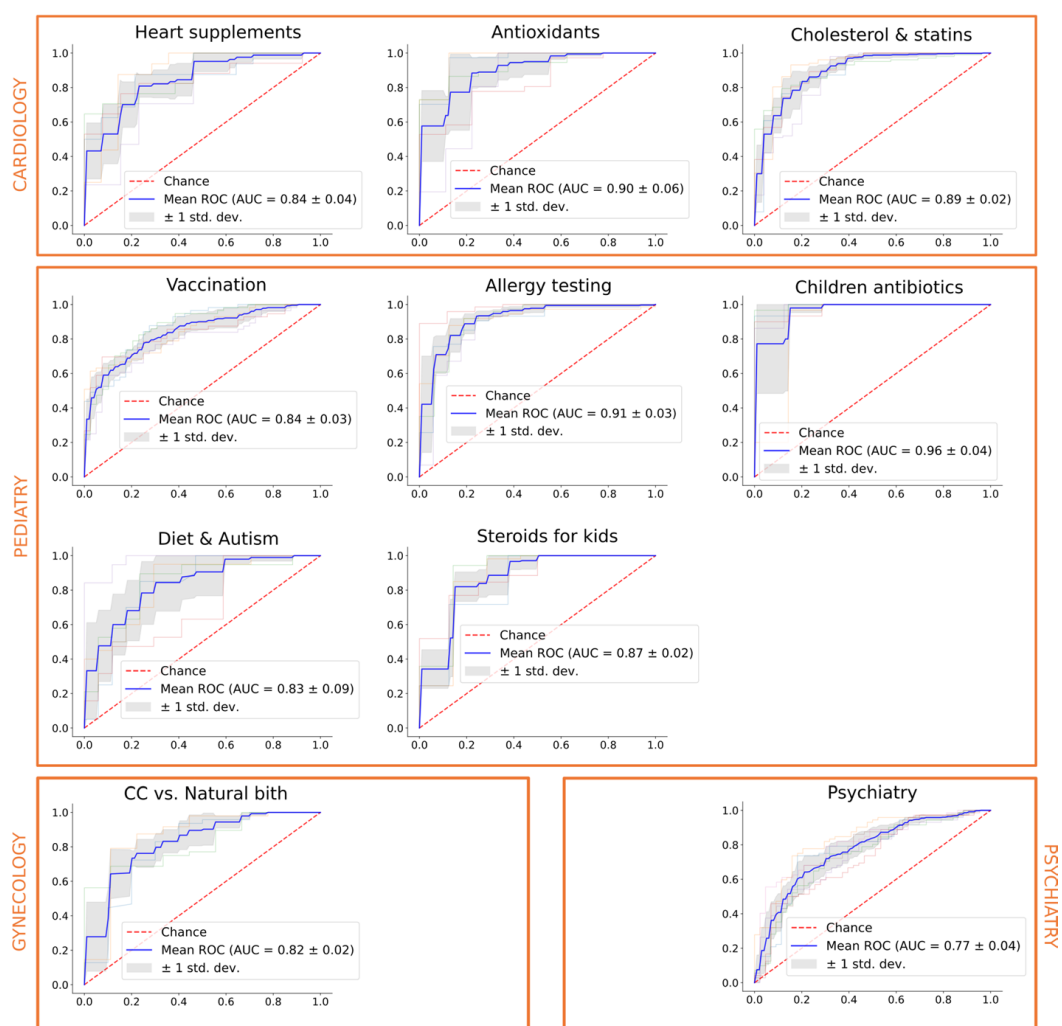


Fig. 1 ROC curves of cross-validated classification results for each medical topic

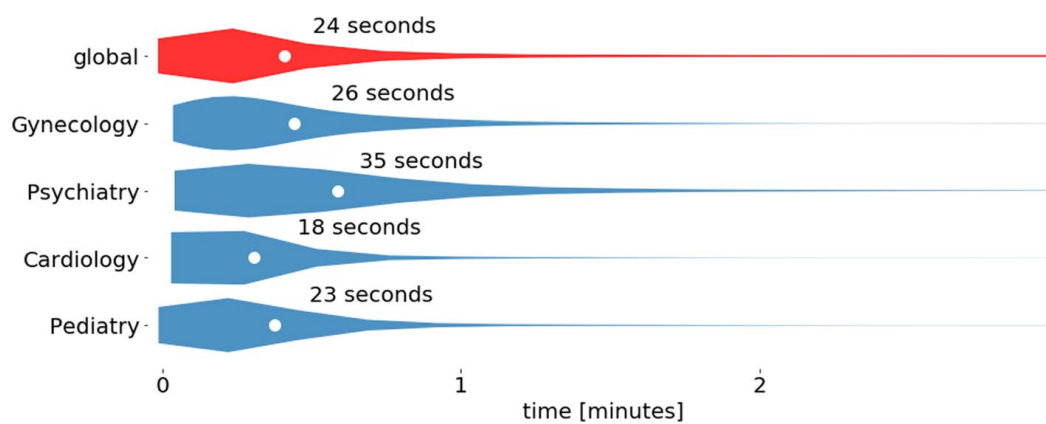


Fig. 2 Times needed to assess a single statement by the medical expert. White dots indicate the average evaluation times, which are explicitly stated in seconds next to each distribution graph

Chapter 5. Articles comprising the thesis

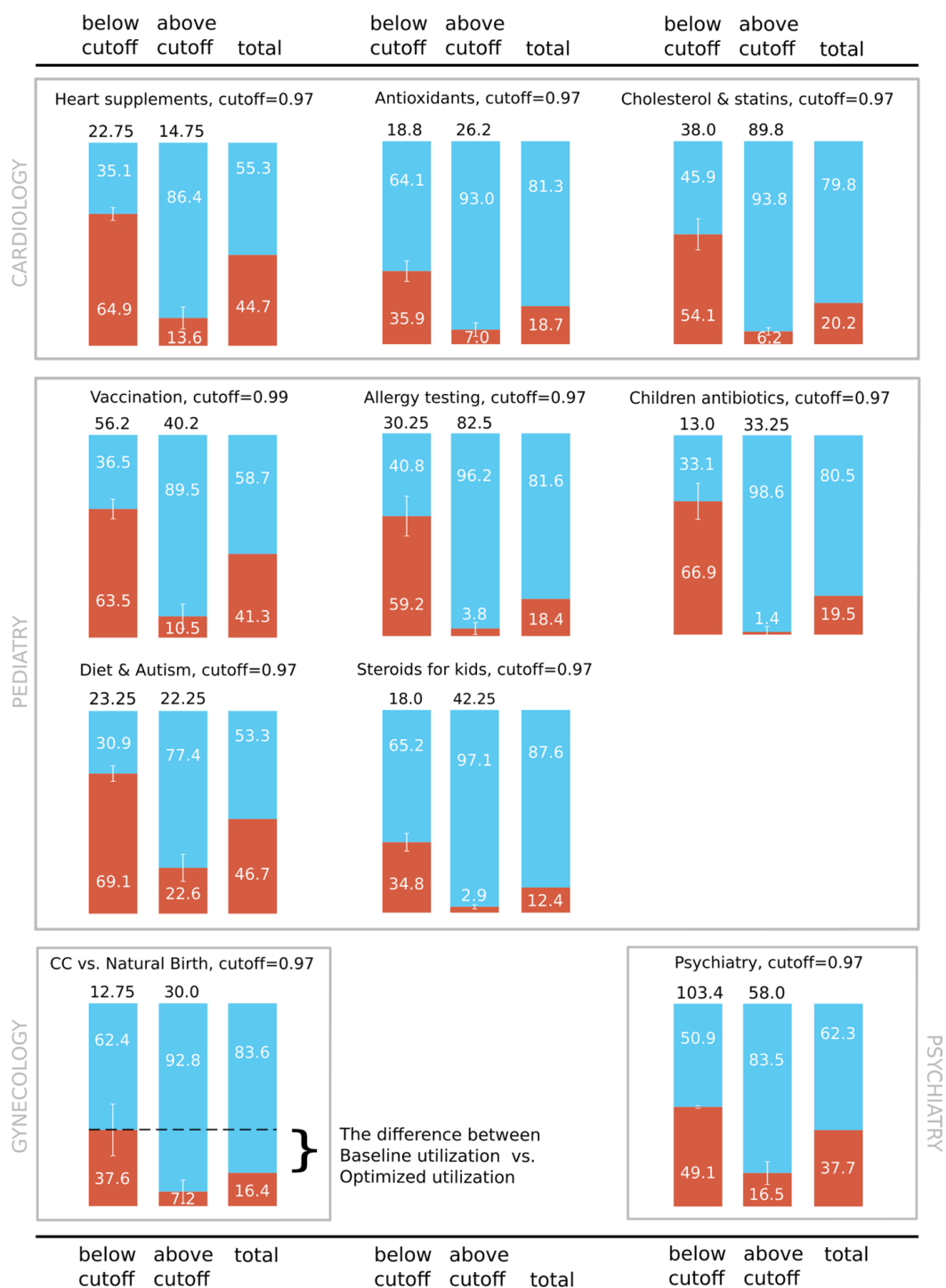


Fig. 3 Cross-validated proportions of positive and negative samples (a) below the cutoff (b) above the cutoff (c) in the entire test set. This corresponds to precision for the negative class, precision for the positive class and total label proportions, respectively. Black labels indicate the mean number of samples in each group. Each bar has the standard deviation indicator (white vertical line)

3.4 Feature selection and model training

The workflow for training statistical models is identical for each topic and includes two steps: feature selection and model selection. Feature selection is performed using Logistic Regression and Recursive Feature Elimination (RFE) [14]. RFE conducts a backward selection of features, starting from a predictive model using all available features. For each feature, the importance score is computed, and the least important feature is removed. The model is retrained with remaining features and the procedure is repeated until the desired number of features remains. We use Logistic Regression as the baseline model for RFE, limiting the number of features to 30% of the number of samples in a given topic. In this paper, we assume that the list of topics is known in advance and that each sentence is already assigned to a topic. This, of course, raises the question of the practical applicability of our method when the topic of an article is unknown. Recent advances in automatic medical subdomain classification [44] suggest that the topic of the article can be successfully extracted from the text.

We have also conducted model training on the unpruned feature set. The results were very disappointing, topical models performed on par with random classification. Thus, we do not include these models in the evaluation. The results for the unpruned feature set strengthen the intuition that credibility assessment is heavily domain-dependent. In our view, this has two consequences. Firstly, the prospects of training a universal credibility assessment model are unlikely as the credibility encoded in the syntax is limited. It seems that most of the credibility is hidden in semantically-loaded features that are specific to a topic. Secondly, the importance of subject matter experts in evaluating the credibility should not be ignored, because only these experts can properly evaluate the significance of topical features. It also stresses the need to augment credibility assessment models with explainability to assist the experts.

For training the model we use the TPOT library [31]. TPOT uses a genetic algorithm to optimize the workflow consisting of feature pre-processing, model selection, and parameter optimization, by evolving a population of workflows and implementing mutation and cross-over operators for workflows. To constrain the space of considered models we use Logistic Regression, XGBoost, and the Multi-layer Perceptron as the initial pool of available models. The optimization is driven by the F_1 measure.

3.5 Explainable models

3.5.1 Models generalization

We try to answer the question about the ability of the models to generalize between sub-domains. To achieve that, we analyzed the most important features for all subdomains with an emphasis on the similarities between the domains (Table 5). We also calculated the percentage of stylometric features from the sets of the most important model features for each sub-domain (Table 4).

3.5.2 Feature weights from logistic regression

All pipelines selected by TPOT involve black-box classifiers and as such cannot be explained globally in terms of feature importance. Only local approximate explanations for

Chapter 5. Articles comprising the thesis

individual samples may be generated by techniques such as SHapley Additive exPlanations (SHAP) [24] or Local Interpretable Model-agnostic Explanations (LIME) [35].

For those subdomains where the F_1 measure and the AUC achieved by Logistic Regression were close to the performance of the pipeline chosen by TPOT (see Appendix A) we used the coefficients of the Logistic Regression models to estimate the importance of each feature and its contribution to the final predictions (see Section 4.4). This may be done since the features were scaled to unit variance.

3.5.3 Locally interpretable model-agnostic explanations

To gain better insight into how filtering classifiers work and boost medical experts' confidence in the robustness of the filtering of credible statements, we perform additional analysis using the locally interpretable model-agnostic explanations (LIME) method [35]. LIME encapsulates any black-box model by a glass-box model (e.g. linear regression or decision tree) operating in the close vicinity of the currently explained instance. The features of the current instance are slightly perturbed (the perturbation type depends on the modality of the instance and may include masking a word or a part of an image, adding noise to the numerical value, flipping of a Boolean value, etc.). The glass-box model is trained only on a small set of perturbations, providing a local approximation of the global (and possibly black-box) model. As the result, the glass-box model identifies features of the explained instance that contribute the most to the current decision of the black-box model.

4 Results

In this section we present the results of conducted experiments. We begin by discussing the process of manual data annotation and its limitations. We show how our active annotation approach optimizes the utilization of subject matter experts' time by re-ranking annotation tasks. We briefly discuss the issue of model generalization, and we conclude the section with extensive analysis of the usefulness of model explainability in credibility assessment.

4.1 Times needed to assess a single statement

During our experiment, we have measured the times required by experts to evaluate the credibility of medical statements. This information is of crucial importance in practice, as the average time to evaluate a statement can be used to determine the throughput of an expert. Of course, it is necessary to keep in mind that experts cannot work indefinitely, and need to take periodic breaks in order to rest.

Figure 2 shows the distributions of evaluation time for all statements, and for statements in the four main disciplines of our study: gynecology, psychiatry, cardiology, and pediatry. The distribution is long-tailed, but the longer times of statement evaluation are infrequent. Overall, the distributions differ for various topics from 18 to 35 seconds, depending on the topic (experts in cardiology are the fastest, while in psychiatry - the slowest). For an expert who works 8 hours per day, with periodic breaks of 15 minutes every hour (leaving 6 hours of effective working time), this gives an average number of evaluated statements per day in the range of 617 to 1200 statements. Recall that, on average, one article in our dataset has approximately 40 statements (there are 10000 statements from 247 articles). This means

that an expert can evaluate from 15 to 30 articles per working day, depending on the topic of the article.

4.2 Optimization of experts' evaluation time

The main objective of our method is to maximize the utilization of medical experts' time when annotating online medical statements. We optimize statistical models to find credible statements, thus increasing the number of non-credible statements that can be presented to medical experts. The results below analyze the efficiency of trained statistical models in finding credible statements. Recall from Section 3.2 that statistical models are trained on a binary dataset consisting of positive (credible and neutral) and negative (non-credible) triplets of sentences.

Figure 1 presents ROC curves for cross-validation. The number of folds depends on the number of samples in a given topic. Based on the ROC curves we have empirically adjusted the cutoff threshold for each classifier's prediction of the positive class. Our goal was to maximize the precision of the negative class while preserving fixed high precision for the positive class. In other words, samples that fall above the cutoff threshold are assumed to contain solely credible or neutral sentences, and will not be presented to medical experts for manual evaluation. We have selected the cutoff threshold for each topic using the following criteria:

- the difference between the proportion of true negative samples and the proportion of negative samples in the entire test set should be maximized, with minimum variance,
- the precision for the true positive class should be maximized,
- the number of samples above the cutoff should be maximized.

The results of the cutoff filtering are presented in Figure 3. For each topic, we show the distribution of positive and negative samples in the entire topic (the *total* column) and in the subsets defined by the cutoff. This corresponds to precision for the negative class (left bar), precision for the positive class (middle bar), and total label proportions (right bar). For instance, there are 44.7% of negative samples and 55.3% of positive samples in the *Heart supplements* topic. The subset of samples defined by the cutoff point of 0.97 contains only 13.6% of negative samples, and the remaining subset contains 64.9% of negative samples. In other words, by removing the samples above the cutoff threshold from manual experts' evaluation we are increasing the number of negative samples that the experts may annotate from 44.7% to 64.9%. We refer to the proportion of negative samples in the topic as the *baseline utilization*, and the proportion of negative samples after the intervention (i.e., below the cutoff threshold) as the *optimized utilization*.

In Table 3 we report baseline utilization, the difference in percentage points with respect to the optimized utilization, and the factor of improvement of medical experts' time utilization. Those values are reported for both models: with TF-IDF and BioBERT lexical features. We denote the percentage point difference value as the *pp. improv.* - percentage point improvement, as for each topic the difference is in favor of using our filtering classifiers.

4.3 Models generalization

Table 4 presents the distribution of significant features between feature sets for TF-IDF and BioBERT-based models. Generally speaking, models built upon TF-IDF vectors are

Chapter 5. Articles comprising the thesis

Table 3 Comparison of baseline and optimized utilization of medical experts’ time

Category	Baseline utilization [%]	pp. improv. [TF-IDF]	factor [TF-IDF]	pp. improv. [BioBERT]	factor [BioBERT]
A	44.7	20.2	1.5	27.6	1.6
B	18.7	17.2	1.9	30.7	2.6
C	20.2	33.9	2.7	21.9	2.1
D	41.3	22.2	1.5	28.1	1.7
E	18.4	40.8	3.2	17.0	1.9
F	19.5	47.4	3.4	35.7	2.8
G	46.7	22.4	1.5	12.2	1.3
H	12.4	22.4	2.8	26.8	3.2
I	16.4	21.2	2.3	25.0	2.5
J	37.7	11.4	1.3	12.9	1.3
Mean	–	25.9	2.2	23.8	2.1

Results presented for both models: (1) using TF-IDF and (2) BioBERT vectors as lexical features. A - heart supplements; B - Antioxidants; C - Cholesterol & statins; D - Vaccination; E - Allergy testing; F - Children antibiotics; G - Diet & Autism; H - Steroids for kids; I - C-section vs. Natural Birth; J - Psychiatry

topic-specific, which may indicate the need for manual fact-checking. However, there are subdomains where the participation of the stylometric features is significant, e.g. *’antioxidants’*. It may be the result of the specificity of this category, where many of the texts were advertisements of either valid or dubious substances.

A much greater share in building filtering classifiers (up to 50% in the case of the category *’heart supplements’*) is when we apply stylometric features along with compressed lexical features, i.e., when the text is embedded using representations extracted from a language model such as BioBERT. Although we lose the ability to directly interpret model decisions related to lexical features (it is not possible to explicitly interpret BioBERT vector’s dimension values), we gain a much greater share of meaningful stylometric features in model construction. There seems to exist a trade-off between lexical and stylometric model explainability, we either explain an algorithmic decision based on lexical features, or based on stylometric features, but not both.

Particularly noteworthy are those stylometric features which have a large share in building filtering classifiers based on BioBERT representations, in particular in the case of categories where models based on BioBERT outperformed models based on TF-IDF. Those models include (per category): statins, antioxidants, vaccination, steroids for kids, C-section vs. natural birth, and (although insignificantly) psychiatry. The features particularly involved in model creation include mostly LIWC features, but also tags retrieved from dependency parsing.

From Table 5 we can see that there are not many stylometric features that are common to all categories (for models built upon TF-IDF vectors). This may indicate that models should be prepared for coherent datasets of very narrow domains.

4.4 Explainable models

For all sub-domains in [Appendix A](#), we present models selected by TPOT. We compare the results of the winning models with the base model, the logistic regression. There are often

Chapter 5. Articles comprising the thesis

Table 4 Percentage of stylometric features from the sets of the most important model features for each sub-domain

Category	LIWC	NER	POS	DEP	Sent	Lexical
TF-IDF						
statins	5.3%	0.0%	0.5%	0.5%	0.5%	93.2%
vaccines	2.8%	0.7%	0.7%	1.4%	0.0%	94.4%
psychiatry	4.3%	0.0%	0.0%	0.0%	0.0%	95.7%
allergy testing	8.2%	0.0%	0.0%	0.0%	0.0%	91.9%
antioxidants	14.7%	0.0%	0.0%	0.0%	0.0%	85.3%
steroids for kids	12.3%	0.0%	0.0%	0.0%	0.0%	87.7%
children antibiotics	3.1%	0.0%	0.0%	0.0%	0.0%	96.9%
diet and autism	5.5%	0.0%	0.0%	0.0%	0.0%	94.5%
heart supplements	12.0%	2.0%	0.0%	0.0%	0.0%	86.0%
cc vs. nb	3.9%	0.0%	0.0%	0.0%	0.0%	96.1%
BioBERT						
statins	12.1%	3.2%	3.2%	4.2%	0.5%	76.8%
vaccines	13.9%	2.9%	2.9%	10.0%	0.7%	69.7%
psychiatry	10.7%	0.0%	2.9%	3.6%	0.0%	82.9%
allergy testing	10.4%	3.00%	2.2%	4.4%	0.0%	80.0%
antioxidants	14.7%	1.4%	0.0%	0.0%	0.0%	84.0%
steroids for kids	21.5%	0.0%	2.8%	8.3%	0.0%	67.4%
children antibiotics	13.9%	3.1%	3.1%	10.8%	0.0%	69.2%
diet and autism	14.6%	1.8%	1.8%	5.5%	0.0%	76.36%
heart supplements	22.0%	8.0%	2.0%	16.0%	2.0%	50.0%
cc vs. nb	22.0%	0.0%	2.0%	14.0%	4.0%	58.0%

LIWC - Linguistic Inquiry Word Count; NER - Named entities count; POS - parts of speech count; DEP - dependency parsing elements count; sent - either polarity or subjectivity of the text; lexical - features that are not stylometric, retrieved either by TF-IDF transformation or the BioBERT model

cases where the logistic regression obtained only slightly worse results than the selected models. For such cases, we assumed that the weights of the logistic regression features are suitable for general explanations of the filtering classifiers' decisions. Feature weight charts for logistic regression for all of the above-defined cases are shown in the [Appendix B](#). Here we present exemplary explanations of models for two topics, antibiotics and dieting in autism, to illustrate the usefulness of having human-interpretable explanations of algorithmic decisions.

4.4.1 Children antibiotics

Figure 4 presents the most important features for distinguishing between credible and non-credible statements regarding the use of antibiotics in children. Features that contribute to the credibility of statements include the use of the word *antibiotic*, the presence of subordinating conjunctions (which characterize complex sentences with constituent subordinate clauses), the presence of "social" vocabulary (i.e., words related to family and friends), as well as the presence of words marking tentative statements (*maybe*, *perhaps*). On the other side, non-credible statements are characterized mostly

Chapter 5. Articles comprising the thesis

Table 5 Number of appearances of those stylometric features that appear more than once per category

Feature name	Number of appearances
Long words (more than 6 letters)	4
Certainty (words such as "always", "never")	3
Emotional tone	2
First person plural count	2
First person singular count	2
Adjectives count	2
Causation (words such as "because", "effect")	2
Past focus (words such as "ago", "did", "talked")	2
Health-related words ("clinic", "flu", "pill")	2
Assent words ("agree", "OK", "yes")	2
Period count	2
Cognitive processes indicators (words such as "cause", "know", "ought")	2
Ingestive processes indicators (words such as "dish", "eat", "pizza")	2

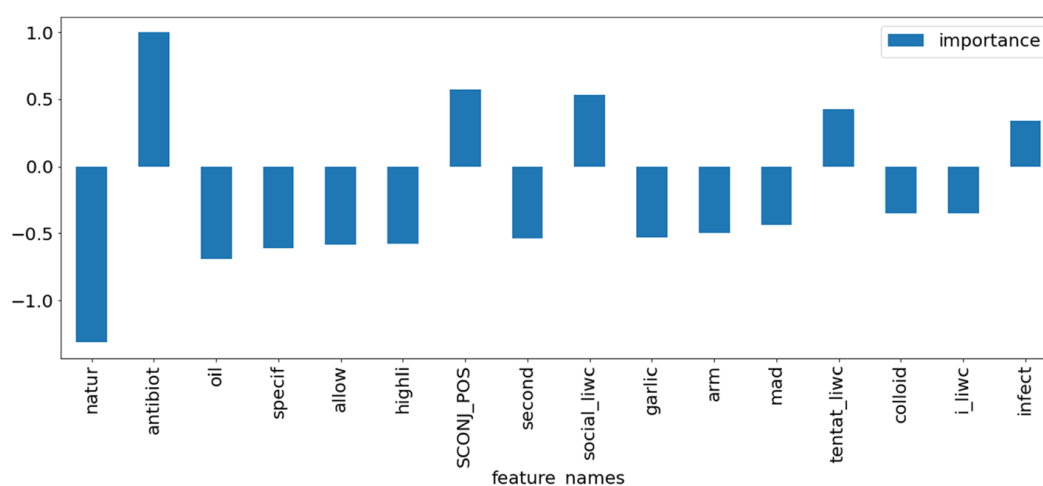
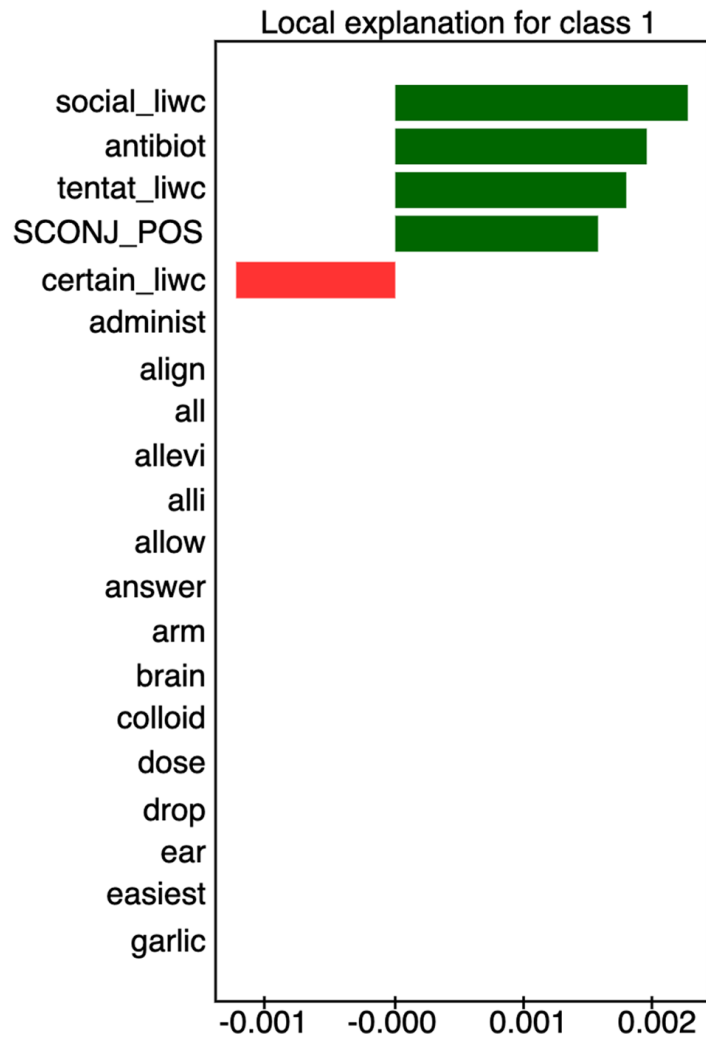


Fig. 4 Feature weights retrieved from Logistic Regression model for 'children antibiotics' category. Top absolute 16 feature weights are depicted

by the presence of specific keywords (*nature, oil, allow, garlic, colloidal silver*). Interestingly, the only keyword marking credible statements is *infection*, which is probably the term avoided by people opposed to the use of antibiotics in children.

Consider the following statement: "However, this study did not determine whether antibiotic use is causally related to breast cancer or if other factors were involved. Certain antibiotics, such as methicillin, vancomycin, sulfonamides, gentamicin, fluoroquinolones, gatifloxacin, levofloxacin, moxifloxacin, and streptomycin, can be harmful for your kidneys. A 2013 study published in the Canadian Medical Association Journal

Fig. 5 LIME explanation for a sentence on antibiotics



found that there is an increase in risk of acute kidney injury among men with use of oral fluoroquinolones.”

This sentence is credible and in line with the current medical knowledge. Figure 5 presents the explanation of the sentence generated by LIME. A medical expert can see that the main reason why this sentence has been classified as credible is the presence of the word *antibiotics* combined with complex phrase structure and tentativeness of the language (*however, whether, did not determine*).

4.4.2 Diet & autism

Most discriminative features for classifying sentences as either credible or non-credible in the domain of diet and autism are depicted in Figure 7. One should remember that this particular subject is extremely sensitive as parents with autistic children may be more vulnerable to exploitation, or easier to accept scientifically unsound recommendations. Features

Chapter 5. Articles comprising the thesis

Fig. 6 LIME explanation for a sentence on diet & autism

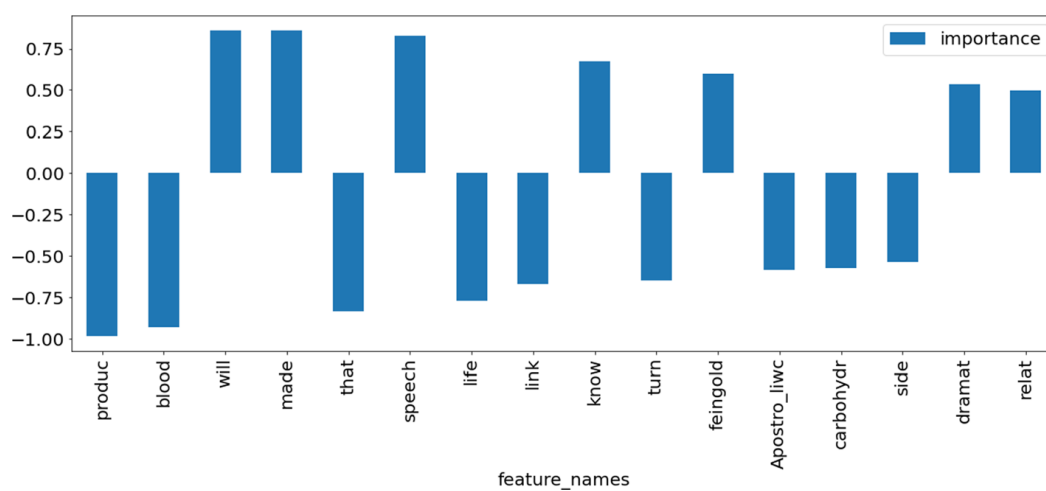
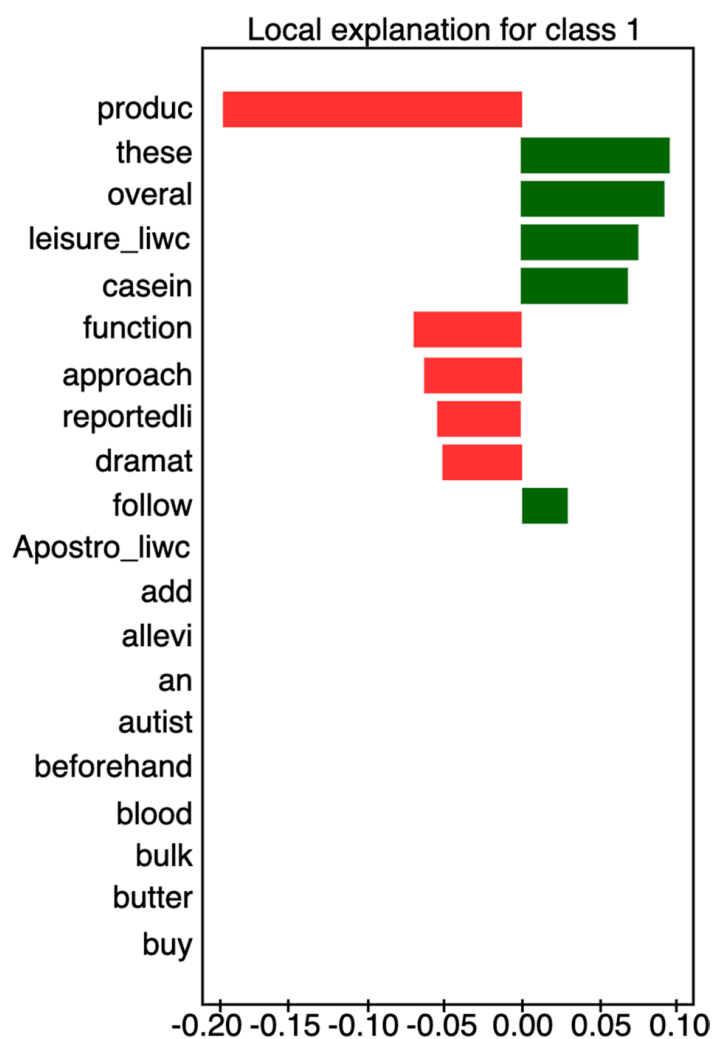


Fig. 7 Feature weights retrieved from Logistic Regression model for 'diet & autism' category. Top absolute 16 feature weights are depicted

Chapter 5. Articles comprising the thesis

characteristic of non-credible statements include very general terms (*product, blood, life, link, turn*) as well as, surprisingly, excessive use of apostrophes. Credible statements also share general terms (*will, made, know, speech, dramatic, relative*), but also mention the Feingold diet, a well-known elimination diet introduced by Benjamin Feingold in the 1970s.

Compare the example of a sentence on antibiotic use with the following non-credible sentence on diet & autism: *"These diets include the following: Casein-free diet (casein is a protein found in milk; this diet eliminates milk and all by-products of milk). In the case of the Autism Spectrum Disorders (ASDs), many parents have reported a reduction in autism symptoms when certain dietary interventions have been tried. For some children, dietary approaches have reportedly produced dramatic changes in overall functioning."*

Figure 6 shows the LIME explanation of the sentence. The sentence is correctly classified as non-credible due to the presence of keywords (*product, function, approach, reported, dramatic*). Keywords associated with credibility (*these, overall, casein*) are not specific enough to sway the decision of the classifier.

5 Discussion

Evaluation of the credibility of online medical information is a very challenging task due to the subjective assessment of credibility, and the specialized medical knowledge required to perform the evaluation [30]. Fully automatic classification of online medical information as credible or non-credible is not a viable solution due to the complex externalities involved in such classification. For the foreseeable future, keeping a human judge in the annotation loop is a necessity. At the same time, qualified human judges are the scarcest resource and their time must be utilized efficiently. Previous approaches to automatically assessing the credibility of medical texts did not take into account the need to weave a human judge into the real-time verification process.

In our work, we present a framework for the optimization of the utilization of medical experts' time when evaluating the credibility of online medical information. To prioritize the evaluation of non-credible information by medical experts, we train classifiers that can filter out credible and neutral medical claims with very high precision exceeding 90% for most medical topics considered in our study (vaccination, allergy testing, children antibiotics, steroids for kids, antioxidants, cholesterol & statins, and C-section vs. natural birth).

Table 3 depicts the key benefit for the potential human-in-the-loop fact-checking system that our solution provides — an increase in the probability that a medical expert will encounter a non-credible medical statement in the annotation batch. As we can see, for all topics the improvement in the utilization of medical experts' time is substantial. The average improvement over all topics is 25.9 percentage points, which means that within the same amount of time and at the same average time needed to annotate a single sentence, medical experts using our method annotate over two times as many non-credible medical statements on average. It is a "pure win" since this improvement does not require any changes to either the annotation protocol or the

annotation interface, we simply make much better use of the experts' time allocated to data annotation.

In addition to the aforementioned important practical implications of using filtering classifiers to prioritize the evaluation of non-credible statements, these classifiers can explain their decisions in a human-interpretable way. Many practical conclusions can be drawn from general and local explanations. For example, the overwhelming share of topic-specific characteristics in classification may indicate that medical fake news are based on certain specific narratives (e.g., vaccines cause autism, high cholesterol is not an indicator of cardiovascular disease) that spread online by copying and pasting or copying and rewriting. This in turn may suggest focusing on semantic similarity measurements as a primary tool for medical fake news detection.

6 Conclusions and future work

One limitation of our method is a certain number of statements that contain misinformation that would not be seen by experts. However, we need to keep in mind that medical experts may not see all statements anyway, as their limited time and attention are not enough to process all suspicious information.

In a realistic use-case scenario, medical experts would continually evaluate a stream of statements derived from the ever-growing set of online articles on medical and health topics, as well as information from social media. Our method increases the efficiency of misinformation detection by medical experts, who will discover more than twice as much misinformation without increasing the time spent on evaluation (or the number of evaluating experts), and without any changes to the annotation workflow. Our method can be regarded as a universal filter for medical Web content. Moreover, we show that we can modify the input features for the filtering classifiers to provide medical experts with different types of feedback, either lexical or stylometric, without any loss of performance. Because we cannot provide medical experts with both lexical and stylometric explanations, it remains to be examined which type of feedback is more useful for medical experts.

In our future work, we plan to focus on gathering more data by introducing the demo expert crowd-sourcing system in selected medical universities. We plan to emphasize the importance of the iterative process of adjusting proper annotation protocol and professional training for medical students. Our goal is to elevate medical students' annotation accuracy to the expert level (like medical practitioners with at least a few years of experience), thus further reducing costs of expert medical credibility annotation.

Appendix A: filtering classifier models

Table 6 presents models selected by TPOT for each subdomain (category), their performance and the comparison to baseline Logistic Regression.

Table 6 Comparison of AUC and weighted F1 performance measures for models selected by TPOT and Logistic Regression for each subdomain

Category	Selected Model (SM)	F1(std) SM	F1(std) LR	AUC(std) SM	AUC(std) LR
A	MLP(50,20) with logistic activation	81(3)	76(5)	87(7)	86(2)
B	MLP(30,20) with ReLU activation	89(2)	80(3)	86(7)	79(6)
C	MLP(50,20) → GradientBoosting → LogisticRegression	77(3)	82(3)	69(5)	81(5)
D	MLP(50,20) with ReLU activation	78(2)	74(5)	84(4)	83(3)
E	MLP(50,20) + MLP(30,20)	87(3)	84(3)	90(4)	79(6)
F	MLP(50,20) with logistic activation	94(2)	96(2)	85(5)	95(5)
G	GradientBoosting → MLP(50, 20) with ReLU activation	73(3)	81(5)	75(5)	80(9)
H	MLP(50,20) with logistic activation	91(1)	71(17)	85(4)	71(3)
I	MLP(50,20) with logistic activation	88(4)	68(8)	83(2)	67(10)
J	GradientBoosting → MLP(50, 20) with ReLU activation	68(4)	71(3)	71(6)	76(5)

A - heart supplements; B - Antioxidants; C - Cholesterol & statins; D - Vaccination; E - Allergy testing; F - Children antibiotics; G - Diet & Autism; H - Steroids for kids; I - CC vs. Natural Birth; J - Psychiatry

Appendix B: logistic regression topical models

Below we present feature weights for Logistic Regression models per each topic.

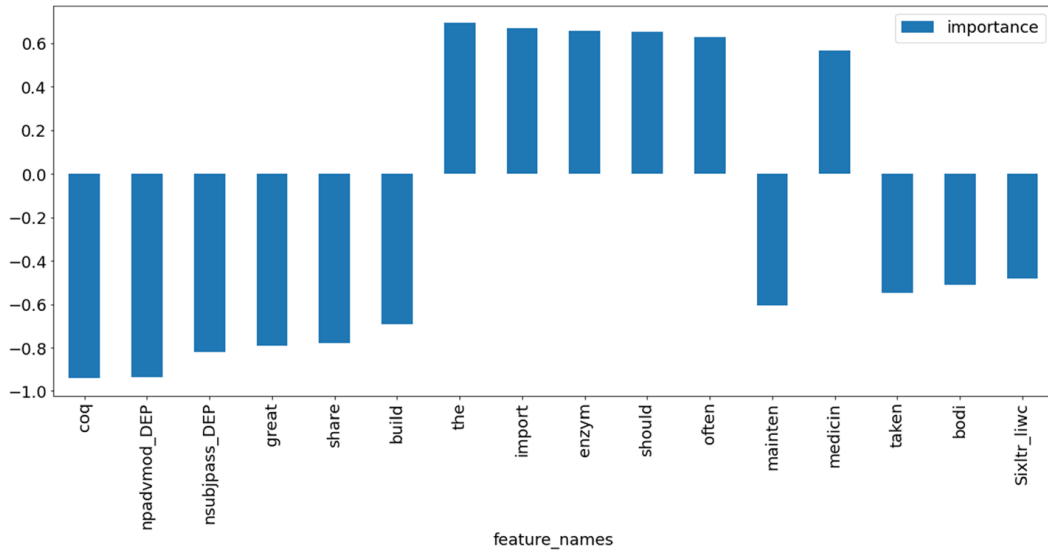


Fig. 8 Feature weights retrieved from Logistic Regression model for 'heart supplements' category. Top absolute 16 feature weights are depicted (roughly 30% of all model features)

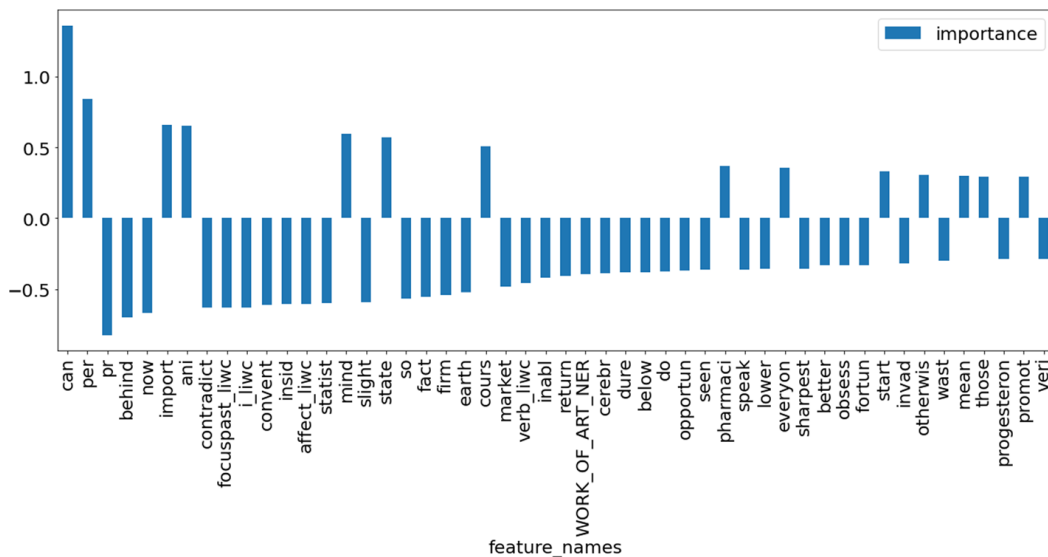


Fig. 9 Feature weights retrieved from Logistic Regression model for 'statins' category. Top absolute 40 feature weights are depicted (roughly 20% of all model features)

Chapter 5. Articles comprising the thesis

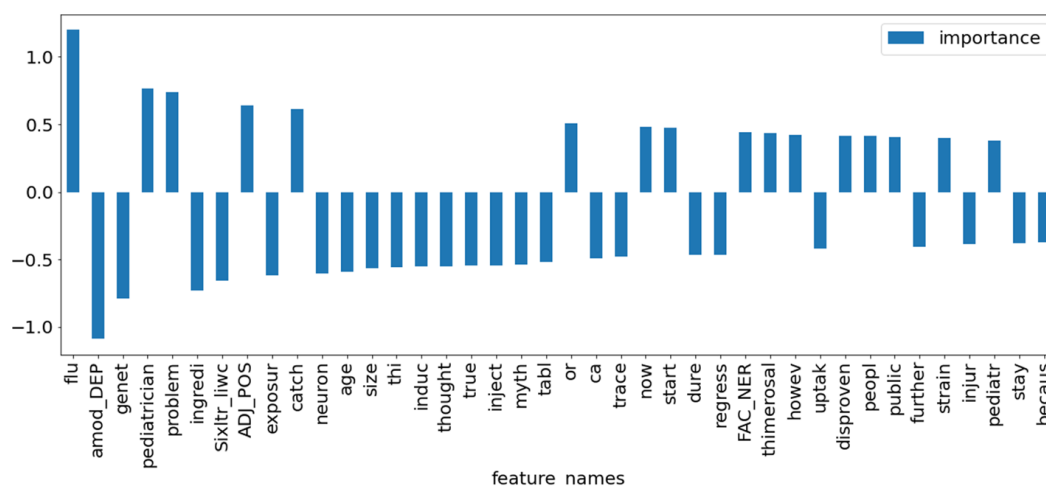


Fig. 10 Feature weights retrieved from Logistic Regression model for 'vaccination' category. Top absolute 40 feature weights are depicted (roughly 30% of all model features)

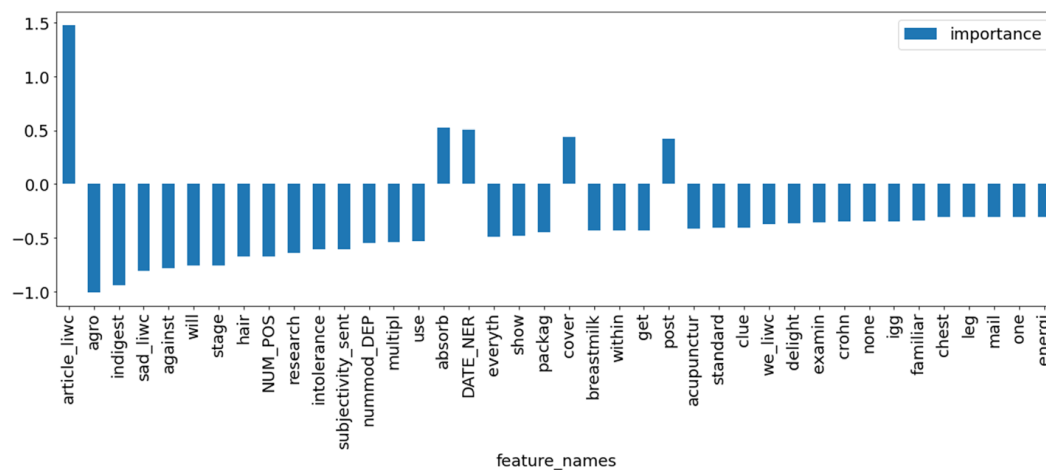


Fig. 11 Feature weights retrieved from Logistic Regression model for 'allergy testing' category. Top absolute 40 feature weights are depicted (roughly 30% of all model features)

Chapter 5. Articles comprising the thesis

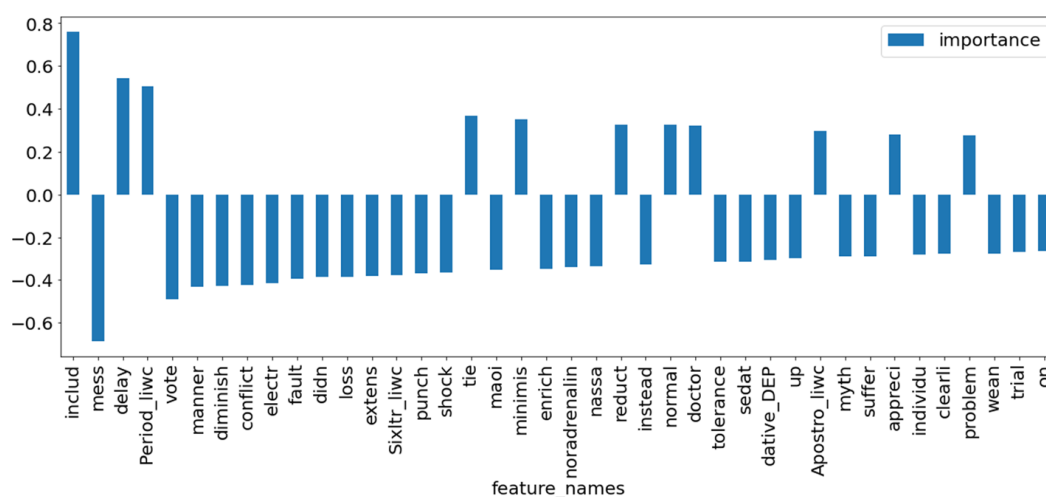


Fig. 12 Feature weights retrieved from Logistic Regression model for 'psychiatry' category. Top absolute 40 feature weights are depicted

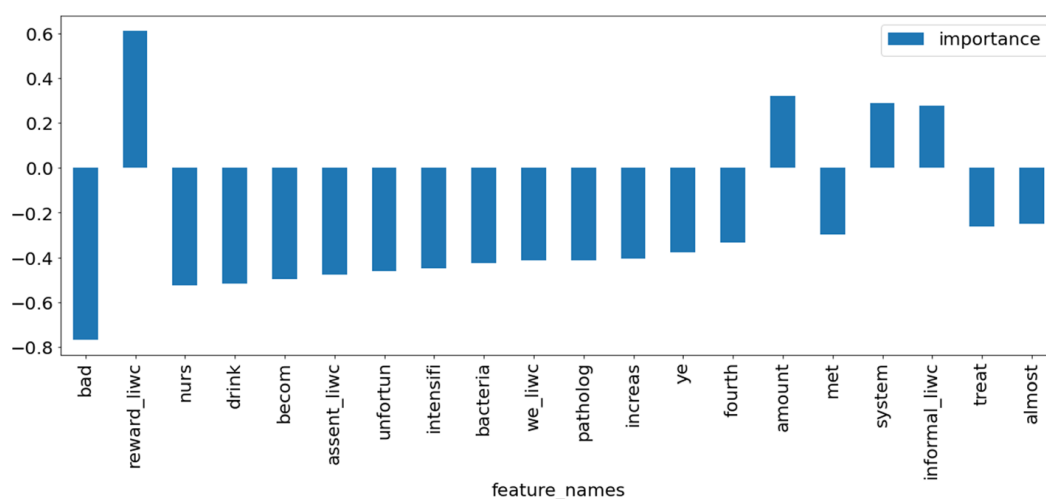


Fig. 13 This data is mandatory. Please check.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Abramczuk, K., Kąkol, M., Wierzbicki, A.: How to support the lay users evaluations of medical information on the Web? https://doi.org/10.1007/978-3-319-40349-6_1 (2016)
2. Afsana, F., Kabir, M A, Hassan, N., Paul, M.: Automatically assessing quality of online health articles. *IEEE J. Biomed. Health Inf.* **25**, 2 (2021). <https://doi.org/10.1109/JBHI.2020.3032479>
3. Balcerzak, B., Jaworski, W., Wierzbicki, A.: Application of textrank algorithm for credibility assessment. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 1, pp 451–454. IEEE (2014)
4. Bode, L., Vraga, E K: See something, say something: Correction of global health misinformation on social media. *Health Commun.* **33**(9), 1131–1140 (2018). <https://doi.org/10.1080/10410236.2017.1331312>
5. Burkart, N., Huber, M F: A survey on the explainability of supervised machine learning. *J. Artif. Intell. Res.* **70**, 245–317 (2021)
6. Chen, Y-Y, Li, C-M, Liang, J-C, Tsai, C-C: Health information obtained from the internet and changes in medical decision making: Questionnaire development and cross-sectional survey. *J. Med. Internet Res.* **20**(2), e47 (2018)
7. Collaboration, S.: Skope-rules. <https://github.com/scikit-learn-contrib/skope-rules>(2020)
8. Davagdorj, K., Park, K H, Amarbayasgalan, T., Munkhdalai, L., Wang, L., Li, M., Ryu, K H: Biobert based efficient clustering framework for biomedical document analysis. In: International Conference on Genetic and Evolutionary Computing, pp 179–188. Springer (2021)
9. Dhoju, S., Main Uddin Rony, M., Ashad Kabir, M., Hassan, N.: Differences in health news from reliable and unreliable media. In: Companion Proceedings of The 2019 World Wide Web Conference. <https://doi.org/10.1145/3308560.3316741>. ACM, New York (2019)
10. Dito, F M, Alqadhi, H A, Alasaadi, A.: Detecting medical rumors on twitter using machine learning. In: 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies, 3ICT 2020. <https://doi.org/10.1109/3ICT51146.2020.9311957>. Institute of Electrical and Electronics Engineers Inc. (2020)
11. Ebnali, M., Kian, C.: Nudge users to healthier decisions: A design approach to encounter misinformation in health forums (2020)
12. Friedman, J H: Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232 (2001)
13. Ghenai, A., Mejova, Y.: Fake cures. *Proc. ACM Human-Comput. Interact.* **2**, CSCW (2018). <https://doi.org/10.1145/3274327>
14. Guyon, I., Weston, J., Barnhill, S.: Gene selection for cancer classification using support vector machines, 46 (2002)
15. Hara, S., Hayashi, K.: Making tree ensembles interpretable. arXiv:1606.05390(2016)
16. Herman, J., Usher, W.: Salib: An open-source python library for sensitivity analysis. *J. Open Source Softw.* **2**(9), 97 (2017)
17. Hou, R., Perez-Rosas, V., Loeb, S., Mihalcea, R.: Towards automatic detection of misinformation in online medical videos. In: 2019 International Conference on Multimodal Interaction. <https://doi.org/10.1145/3340555.3353763>. ACM, New York (2019)
18. Jensen, M L, Averbeck, J M, Zhang, Z., Wright, K B: Credibility of anonymous online product reviews: A language expectancy perspective. *J. Manag. Inf. Syst.* **30**, 1 (2013). <https://doi.org/10.2753/MIS0742-1222300109>
19. Latkin, C A, Dayton, L., Yi, G., Konstantopoulos, A., Boodram, B.: Trust in a COVID-19 vaccine in the U.S.: A social-ecological perspective. *Social Science & Medicine*, 270. <https://doi.org/10.1016/j.socscimed.2021.113684> (2021)
20. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C H, Kang, J.: Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4), 1234–1240 (2020)
21. Li, J.: Detecting false information in medical and healthcare domains: A text mining approach. https://doi.org/10.1007/978-3-030-34482-5_21 (2019)
22. Liu, X., Zhang, B., Susarla, A., Padman, R.: YouTube for patient education: A deep learning approach for understanding medical knowledge from user-generated videos. *ArXiv Computer Science* (20187)
23. Lou, Y., Caruana, R., Gehrke, J., Hooker, G.: Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 623–631 (2013)
24. Lundberg, S M, Lee, S-I: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp 4768–4777 (2017)
25. Molnar, C.: *Interpretable Machine Learning*. Lulu.com (2020)
26. Morris, M D: Factorial sampling plans for preliminary computational experiments. *Technometrics* **33**(2), 161–174 (1991)

Chapter 5. Articles comprising the thesis

27. Murdoch, W J, Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.* **116**(44), 22071–22080 (2019)
28. Nabożny, A, Balcerzak, B., Morzy, M., Wierzbicki, A.: Focus on misinformation: Improving medical experts' efficiency of misinformation detection. In: *International Conference on Web Information Systems Engineering*, pp 420–434. Springer (2021)
29. Nabożny, A, Balcerzak, B., Wierzbicki, A., Morzy, M., Chlabicz, M.: Active annotation in evaluating the credibility of Web-based medical information: Guidelines for creating training data sets for machine learning. *JMIR Med. Inform* **9**(11), e26065 (2021). <https://doi.org/10.2196/26065>, <https://medinform.jmir.org/2021/11/e26065>
30. Nabożny, A, Balcerzak, B., Wierzbicki, A., Morzy, M., Chlabicz, M., et al.: Active annotation in evaluating the credibility of Web-based medical information: Guidelines for creating training data sets for machine learning. *JMIR Med. Inf.* **9**(11), e26065 (2021)
31. Olson, R S, Urbanowicz, R J, Andrews, P C, Lavender, N A, Kidd, L C, Moore, J H: Automating biomedical data science through tree-based pipeline optimization. <https://epistasislab.github.io/tpot/citing/> (2016)
32. Pollard, M S, Davis, L.M.: Decline in trust in the centers for disease control and prevention during the COVID-19 pandemic. <https://doi.org/10.7249/RRA308-12> (2021)
33. Purnomo, M H, Sumpeno, S., Setiawan, E I, Purwitasari, D.: Biomedical engineering research in the social network analysis era: Stance classification for analysis of hoax medical news in social media. *Procedia Computer Science*, 116. <https://doi.org/10.1016/j.procs.2017.10.049> (2017)
34. Rafalak, M., Abramczuk, K., Wierzbicki, A.: Incredible: Is (almost) all Web content trustworthy? Analysis of psychological factors related to website credibility evaluation. In: *Proceedings of the 23rd International Conference on World Wide Web*, pp 1117–1122 (2014)
35. Ribeiro, M T, Singh, S., Guestrin, C.: “why should i trust you?” Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 1135–1144 (2016)
36. Samory, M., Mitra, T.: ‘The government spies using our webcams’: The language of conspiracy theories in online discussions. *Proceedings of the ACM on Human-Computer Interaction*, **2**(CSCW). <https://doi.org/10.1145/3274421> (2018)
37. Samuel, H., Zaïane, O: *MedFact: Towards improving veracity of medical information in social media using applied machine learning* (2018)
38. Sicilia, R., Lo Giudice, S., Pei, Y., Pechenizkiy, M., Soda, P.: Twitter rumour detection in the health domain. *Expert Syst. Appl.*, 110. <https://doi.org/10.1016/j.eswa.2018.05.019> (2018)
39. Singh, C., Nasser, K., Tan, Y S, Tang, T., Yu, B.: imodels: A python package for fitting interpretable models. *Open J* **6**, 61 (2021). <https://doi.org/10.21105/joss.03192>
40. Wagle, V., Kaur, K., Kamat, P., Patil, S., Kotecha, K.: Explainable ai for multimodal credibility analysis: Case study of online beauty health (mis)-information. *IEEE Access* **9**, 127985–128022 (2021)
41. Walter, N., Brooks, J J, Saucier, C J, Suresh, S.: Evaluating the impact of attempts to correct health misinformation on social media: A meta-analysis. *Health Commun.* <https://doi.org/10.1080/10410236.2020.1794553> (2020)
42. Wang, Y., McKee, M., Torbica, A., Stuckler, D.: Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, 240. <https://doi.org/10.1016/j.socscimed.2019.112552> (2019)
43. Wang, Z., Yin, Z., Argyris, Y A: Detecting medical misinformation on social media using multimodal deep learning. *arXiv* (2020)
44. Weng, W-H, Waghlikar, K B, McCray, A T, Szolovits, P., Chueh, H C: Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Med. Inform. Decis. Making* **17**(1), 1–13 (2017)
45. Wierzbicki, A.: *Web Content Credibility*. Springer (2018)
46. Xu, Z., Guo, H.: Using text mining to compare online pro- and anti-vaccine headlines: Word usage, sentiments, and online popularity. *Commun. Stud.* **69**(1), 103–122 (2018). <https://doi.org/10.1080/10510974.2017.1414068>
47. Zhang, X., Ghorbani, A A: An overview of online fake news: Characterization, detection, and discussion. *Inf. Process. Manag.* **57**, 2 (2020). <https://doi.org/10.1016/j.ipm.2019.03.004>
48. 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 (2021). (<https://doi.org/10.1016/j.ipm.2020.102390>)
49. Zhu, Y., Li, L., Lu, H., Zhou, A., Qin, X.: Extracting drug-drug interactions from texts with biobert and multiple entity-aware attentions. *J. Biomed. Inform.* **106**, 103451 (2020)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.