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-

sessment. 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.