


TA-WHI: Text Analysis of Web-Based Health Information

Piyush Bagla, Dr. B.R. Ambedkar National Institute of Technology, Jalandhar, India*

 <https://orcid.org/0000-0002-1664-7787>

Kuldeep Kumar, National Institute of Technology Kurukshetra, India

ABSTRACT

The healthcare data available on social media has exploded in recent years. The cures and treatments suggested by non-medical experts can lead to more damage than expected. Assuring the credibility of the information conveyed is an enormous challenge. This study aims to categorize the credibility of online health information into multiple classes. This paper proposes a model named Text Analysis of Web-based Health Information (TA-WHI), based on an algorithm designed for this. It categorizes health-related social media feeds into five categories: sufficient, fabricated, meaningful, advertisement, and misleading. The authors have created their own labeled dataset for this model. For data cleaning, they have designed a dictionary having nouns, adverbs, adjectives, negative words, positive words, and medical terms named MeDF. Using polarity and conditional procedure, the data is ranked and classified into multiple classes. The authors evaluate the performance of the model using deep-learning classifiers such as CNN, LSTM, and CatBoost. The suggested model has attained an accuracy of 98% with CatBoost.

KEYWORDS

Credibility, Data Mining, Health information, Machine Learning, NLP, Online Health, Text Mining, WHI

INTRODUCTION

In the healthcare industry, wrong treatment, misinformation, self-treatment, and myths related to unconventional treatments is not a recent development. It is as ancient as medical care itself. Before the boom of the Internet, Radio, and Television, this issue was based on the therapeutic relationship as well as its context (Fernández-Celemín & Jung, 2006). The spectrum of damage is taken to an entirely new degree because of global technological advancement. Misinformation on social media became so common that in 2016 Oxford dictionary introduced “post-truth,” meaning “relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief” (Harsin, 2018). Posting misleading or misinformation on social media is a fashion for some.

Social networking services like Facebook, followed by Twitter, are currently the industry leaders, with over 1.3 billion members and a monthly average fluctuation of 300 million people. Every second,

DOI: 10.4018/IJSSCI.316972

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

their interactions create gigabytes of data (Alrubaian et al., 2018; Ranganath et al., 2017). Online social networks are appealing because they provide a quick and easy way to acquire health information. It is also quite simple to share information with others. However, the broad dissemination of incorrect information is made possible by rapid data scattering at a high pace with little effort. Thanks to the pandemic in 2020, social media usage increased by many folds. More information is now shared on social media than before 2020 (Zhang et al., 2017). The world has seen how misinformation about COVID spreads like wildfire, and every time World Health Organization (WHO) or some medical authority comes up to deny the news. People are so scared to visit hospitals that they prefer the social media Doctor (Zwolenski & Weatherill, 2014). Following the incorrect therapeutic advice given on social media might be fatal.

Text analysis is the practice of analyzing a vast amount of textual material to capture the key concept, trends, and hidden relationships. It is the process of transforming the unstructured text into a structured format to identify meaning patterns and new insights. Analysis of text is a crucial step in getting the hidden meaning behind it. The most popular technique for doing so is sentiment analysis. There are a number of researchers who have used this technique to get the actual sentiment behind the post on social media, especially on Twitter. It is extended further to incorporate a machine learning algorithm to perform the classification task. As a result, the credibility of the post can be identified (Alharbi & Alhalabi, 2020; Gunti et al., 2022; Mohammed et al., 2022). People are now mixing their regional language while making any social media post, no matter from which country they belong. In technical terms, we call this code-mixing. This makes the analysis of text even more difficult. However, techniques such as Machine Learning, Neural Networks, and LSTM (Long Short-Term Memory) can be used to mitigate the problem of code-mixing (Sharma et al., 2021; Singh & Sachan, 2021). With the development of technology, the amount of data generated on the Internet has increased daily. This data includes valuable patterns that must be recognized to get meaningful information. There are several methods that facilitate the completion of this task, such as using data mining techniques (García-Peñalvo et al., 2021), text mining and privacy preservation techniques in name analysis (Veluru et al., 2015), and scientific issue tracking with topic analysis based on crowdsourcing (Kim et al., 2018). In one way or another, all these methods contribute to the data mining process. However, there is always room for strengthening the capabilities of the proposed approaches. For example, there is always a question regarding the authenticity of the information patterns found during text mining. Very few studies address this issue and those that do have several drawbacks. To determine the veracity of web-based health information, they used a predetermined data mining algorithm that operates on the existing dataset. However, there is an ongoing need to develop a strategy based on a user-generated algorithmic approach to determining the credibility of web-based health information using real-time datasets.

We may break up research on automatic reliability evaluation for online news into three sections. In knowledge bases, some people try to extract and validate allegations stated in the text. It is further known as “fact-check.” Others focus on measuring trustworthiness in the news context content’s social media environment and its creator (Atanasova et al., 2019; Ciampaglia, 2015; Thorne et al., 2018). The third option depends on the text’s general linguistic characteristics, such as writing style. Low-credibility material, such as fake news, is written to evoke an intense emotional response (Tacchini et al., 2017; Zubiaga et al., 2016), which requires specific stylistic methods. Measurement of linguistic complexity, detection of syntactic patterns using n-grams of part-of-speech tags or counting words belonging to particular categories might act as credibility indicators (Ahmed et al., 2021; Bhattarai et al., 2021; Potthast et al., 2017; Reis et al., 2019).

The remaining sections are organized as follows: First, we explore related work, including some of the most important research undertaken in this field. We then explain our proposed methodology, which is further broken into many sub-sections. The study’s results are presented in the next section, followed by the conclusion.

12 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/ta-whi/316972

Related Content

Chaotic Tornadoogenesis Optimization Algorithm for Data Clustering Problems

Ravi Kumar Saidalaand Nagaraju Devarakonda (2018). *International Journal of Software Science and Computational Intelligence* (pp. 38-64).

www.irma-international.org/article/chaotic-tornadoogenesis-optimization-algorithm-for-data-clustering-problems/199016

A Bayesian Based Machine Learning Application to Task Analysis

Shu-Chiang Lin (2012). *Machine Learning: Concepts, Methodologies, Tools and Applications* (pp. 234-242).

www.irma-international.org/chapter/bayesian-based-machine-learning-application/56143

Improving Accuracy of Event-Related Potentials Classification by Channel Selection Using Independent Component Analysis and Least Square Methods

Wenxuan Li, Mengfan Liand Wei Li (2016). *International Journal of Software Science and Computational Intelligence* (pp. 1-18).

www.irma-international.org/article/improving-accuracy-of-event-related-potentials-classification-by-channel-selection-using-independent-component-analysis-and-least-square-methods/172124

A Novel Chaotic Northern Bald Ibis Optimization Algorithm for Solving Different Cluster Problems [ICCICC18 #155]

Ravi Kumar Saidalaand Nagaraju Devarakonda (2019). *International Journal of Software Science and Computational Intelligence* (pp. 1-25).

www.irma-international.org/article/a-novel-chaotic-northern-bald-ibis-optimization-algorithm-for-solving-different-cluster-problems-iccicc18-155/233520

Cognitive Processes of the Elderly Brain with MindGym Approach

Jurij Tasic, Darja Rudan Tasic, Andrej Vovk, Dusan Šuput, Shushma Patel, Dilip Patel, Marjan Gusevand Sasko Ristov (2015). *International Journal of Software Science and Computational Intelligence* (pp. 18-34).

www.irma-international.org/article/cognitive-processes-of-the-elderly-brain-with-mindgym-approach/157435