Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 220 (2023) 820–825

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Workshop on Hospital 4.0 (Hospital)
March 15 - 17, 2023, Leuven, Belgium

An Architecture Proposal for Noncommunicable Diseases Prevention

Diana Braga[a], Daniela Oliveira[a], Rafaela Rosário[b], Paulo Novais[a], José Machado[a,*]

[a]*ALGORITMI Research Center, University of Minho, 4800-058 Guimarães, Portugal*
[b]*School of Nursing, University of Minho, 4710-057 Braga, Portugal*

Abstract

Noncommunicable Diseases (NCDs) are a leading global health challenge, causing 41 million deaths per year. Risk factors include genetics, environmental factors, and lifestyle choices. Adopting healthy lifestyles can prevent or delay the onset of NCDs, but health misinformation can lead people to make poor decisions about their health.

To combat this, it is proposed to develop an Intelligent System using Artificial Intelligence techniques to collect and analyze data from social media about health topics to combat misinformation in public health and forecast NCDs, providing guidelines to prevent their spread. Methods: A system´s overall architecture is presented ... An innovative and novel solution that addresses the spread of information concerning health and NCDs contributes to inform public policies and infodemic management strategies

Keywords: Intelligent System; Misinformation; Noncommunicable Diseases; Prevention; Online Social Listening

1. Introduction

Noncommunicable Diseases (NCDs), also known as chronic diseases, are long-term illnesses caused by a combination of genetic, physiological, environmental, and behavioral factors. Cardiovascular diseases, cancers, diabetes, and chronic respiratory diseases are the most common types of NCDs and they are responsible for a large proportion of deaths worldwide, accounting for 74% of all deaths globally [1].

This is a serious global health challenge, driven by factors such as growing unplanned urbanization, unhealthy lifestyles, and population aging [2]. However, most of these risk factors can be modified or controlled through different strategies, and if they are minimized, it is possible to prevent 80% of all heart disease, stroke, diabetes, and 40% of malignancies [3]. It is particularly important to prevent NCDs in children and adolescents because the risks of these

* Corresponding author.
E-mail address: jmac@di.uminho.pt

1877-0509
10.1016/j.procs.2023.03.109

conditions often develop in childhood and can persist into adulthood, increasing the burden for public health and health systems. Strategies to prevent NCDs in this age group include promoting healthy lifestyles, such as a healthy diet and healthy 24h movement behavior (e.g., physical activity, sedentary behavior and sleep), through education, policy, and environmental changes [4].

Social media usage is rising steadily over the world, including in situations related to health issues [5], since it has resulted in an easily accessible supply of real-time data about what users say and share in online groups [6]. For this reason, people are increasingly looking for health information from digital sources; however, despite its potential as an ally in obtaining accurate information, the presence of mis and disinformation on social media is a major concern, since it can lead people to make poor health decisions [7, 8]. In fact, false information has been demonstrated to be published twice as frequently as evidence-based information during health emergencies, which generates more panic and doubts in the population [6]. The COVID-19 pandemic was the most dramatic demonstration of the significance of fighting mis and disinformation, where it was possible to show how pervasive and damaging it can be [9].

To combat this problem, there has been an increasing focus on infodemic management systems, which are designed to counter the negative effects of spreading misinformation, by promoting accurate and reliable information and reducing its spread [10]. This can be achieved through various strategies, like having a central source of accurate information, distributing it via multiple channels, identifying and correcting misinformation, developing strategies to prevent its spread, and engaging key stakeholders to promote accurate information and counter misinformation [7]. To build such a system for the prevention of NCDs is important to understand how health misinformation and NCDs risk factors are related, so combating mis and disinformation about healthy habits and lifestyles can help control these risk factors and minimize the burden of NCDs on individuals and society, especially in the younger age group.

2. Related Work

DataReportal [11] reports that the worldwide mobile population has surpassed 5 billion users in October 2022, and Li et al. [12] shows that over 70% of individuals use the internet to seek healthcare-related information. However, this poses a significant health problem as people may easily be misinformed due to the abundance of false information available online [13]. For these reasons, the identification of health misinformation in social media has piqued the curiosity of several researchers during the last decade. This section covers several research that focuses on online social listening to prevent the infodemic in health.

A research team from a college in Lisbon developed the CovidCheck.pt website [14], a project funded by the *Gulbenkian Soluções Digitais Covid-19* initiative. Its goal is to improve official communication and address the main concerns of the Portuguese population about the Covid-19 pandemic, by providing daily updates since May 11, 2020. This tool mainly aims to identify misinformation that could harm public health and encourage the public to seek reliable sources. The study analyzes four different aspects of public discourse: official discourse topics, media information, the public expression on social media (concerning doubts, irresponsible or dishonest behavior, and misinformation), and public concerns expressed in web searches. After being gathered, the data is encoded, categorised, and confirmed by psychologists before being converted into suggestions for improving communication with citizens, the elderly, and other high-risk populations. The site has different sections, one with recommendations that focus on answering questions and doubts about COVID-19 such as ”can risk groups use homemade masks?” and another section called ”Don’t be Misled” which is designed to combat misinformation by identifying and correcting false information related to COVID-19 that circulates in social media.

Klein et al. [15] published a study in which they used a Natural Language Processing (NLP) pipeline to collect data from Twitter in order to identify potential cases of COVID-19 in the United States that were not based on testing. They collected English tweets containing keywords related to COVID-19 and applied regular expressions to identify tweets that suggested the user may have been exposed to the virus. They then trained and evaluated two Deep Neural Network (DNN) classifiers based on BERT, BERT-Base-Uncased and COVID-Twitter-BERT, on a sample of the tweets. COVID-Twitter-BERT achieved an F1-score of 0.76 for detecting tweets that self-reported potential cases of COVID-19. They then applied the automatic pipeline to over 85 million unlabeled tweets collected between March and August 2020, identifying 13,714 tweets that self-reported potential cases of COVID-19 with US state-level geolocations.

The World Health Organization (WHO) and its research partners developed a pilot project called Early Artificial intelligence-supported Response with Social Listening (EARS) in 20 countries [16]. The project aimed to use social listening and AI to identify and evaluate public opinions and concerns related to health through a taxonomy. The EARS project was designed to help detect and respond to potential public health emergencies by quickly analyzing data from various sources. WHO is collecting data daily from online conversations in publicly available sources, including Twitter, Facebook public pages, online forums, news comments, and blogs in 30 pilot countries in nine languages. The data collected from these sources is normalized and sampled to make it usable and in order to make comparisons between countries with different population sizes and internet access levels fair. The data is collected by using a query that includes broad COVID-19 keywords, and this is done to control the amount of data that is processed. Although the project allows for country-level and cross-country comparisons of online conversations, a human analyst is still needed to identify potential gaps in information and causes of misunderstanding. The collected data is categorized into defined topics of interest by health information experts and through an analysis of the data. The data is automatically categorized through semi-supervised Machine Learning (ML) and human quality controls, and the system adapts to language and cultural differences in each country. Additionally, opinions can be filtered by intent, such as questions or complaints, which are automatically detected by the system.

Purushothaman et al. in their study [18] aimed to analyze the characteristics of content related to nicotine poisoning on the popular social media platform TikTok. They collected posts associated with the hashtag #nicsick using a Python package (SeleniumHQ), analyzed the content and characteristics of the videos, and analyzed metrics such as user engagement, video characteristics, and video type. They found that over half of the videos discussed firsthand and secondhand reports of suspected nicotine poisoning symptoms and experiences. They suggested that digital surveillance on social media platforms like TikTok could be used to detect vaping-related adverse events, particularly among young people.

Overall, social listening on online platforms has been demonstrated to be a useful method for understanding public health concerns, knowledge, and behaviors. By analyzing conversations and content on social media, important insights can be gained about various health-related topics.

3. Architecture Proposal

Given the context described above, the primary goal of this research is to purpose a solution based on the development of an Intelligent System (IS) for preventing NCDs in society by building a social sensing data collection platform that can gather relevant information from online social media users about lifestyles and health-related topics. This information will be used to identify potential risk factors for NCDs, such as unhealthy diets, lack of physical activity, and smoking.

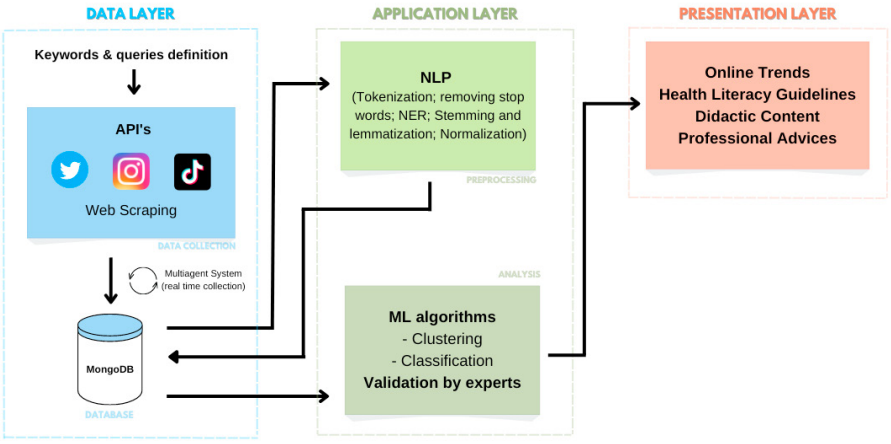

Fig. 1. Social Media Monitoring and Analysis Pipeline.

The system's overall architecture, as depicted in Figure 1, illustrates the various stages of development, and it's based on the Three-Tier Architecture methodology, which divides computing into three logical and physical tiers: the data tier, where application data is stored and managed; the application tier, where data is processed; and the presentation tier.

1. Data Layer

- **Define the research question** and identify keywords related to NCDs. It is important to search for information not only about NCDs themselves but also about the **risk factors** associated with them, such as habits and healthy lifestyles, diets, access to healthy food, physical exercise, sleep quality, alcohol, and smoking in order to fully understand the impact on individuals and communities. The keywords related to **"children and teenagers"** are particularly important in this case as it pertains to understanding and addressing the potential impact of misinformation on the health and well-being of future generations.
- **Collect data** from social media APIs such as Twitter, Instagram, TikTok, and other online sources of health information (such as news APIs). In the specific case of TwitterAPI, it is important to collect not only the tweets but also user information, tweet engagement (likes, retweets), geolocation, URLs cited, and other relevant data. Web scraping techniques can be useful to gather information from news sources.
- **Multi-Agent System (MAS)** to control the data collection process. Select the attributes of interest, remove duplicate data, and export the cleaned data to a **MongoDB database**. This phase consists to upload a dataset with attributes already defined in a global way, to facilitate the subsequent cleaning of the data. It can however be done in parallel with that phase.

2. Application Layer

- **Preprocess the data** by cleaning and extracting information using **NLP** techniques. For example, tokenization can be used to break down the text into individual words or phrases, while stemming or lemmatization can be used to standardize the word forms. Additionally, techniques such as sentiment analysis, entity recognition, and named entity recognition (NER) can be applied to identify the polarity, subject, and entities in the text respectively. Text classification techniques such as bag-of-words or word embeddings can also be utilized to classify text data into different categories, such as the source of the data, the credibility of the source, etc.
- Store the clean data in a new MongoDB collection or in a file-based **Big Data architecture**.
- **Analyze the data** to evaluate its accuracy and reliability. ML algorithms will be used to classify and cluster the information, in order to determine whether it is valid or not. Furthermore, it can be useful to use input from health professionals, such as nurses and nutritionists, to evaluate specific information and to identify misinformation.

3. Presentation Layer

- **Online Trends** presented in the platform that will be developed to monitor and analyze social media conversations in real-time to identify trends and patterns related to health topics. This functionality will allow users to stay informed about the latest health trends and conversations taking place on social media. It can be useful for public health professionals and policymakers to identify emerging health concerns and take appropriate action, but also for individual users who want to be aware of current health issues.
- **Health Literacy Guidelines** provided to the users with clear and easy-to-understand health information that can help them make informed decisions about their health. The information will be based on scientific evidence and guidelines and will be regularly updated to ensure it is accurate and up-to-date.
- **Didactic Content** to atract different type of users with engaging and interactive didactic content such as videos that make learning about health topics fun and easy, particularly for children.
- **Professional Advices** from healthcare professionals, such as nurses, psychologists, and nutritionists. The professional advice will be based on general health issues and good lifestyle habits that can prevent the appearance of diseases in the future.

Overall, the goal of this research is to develop an IS that can automatically detect potential health crises and provide targeted interventions to mitigate their impact on public health by using social sensing data collection. The IS could be a valuable tool for public health professionals and policymakers in their efforts to prevent NCDs in society.

4. Conclusion and Future Work

The proposed project addresses a critical problem in public health: mis and desinformation about health topics and NCDs. The proposed solution is an innovative approach that uses AI and NLP techniques to extract relevant information from social media and provide accurate and up-to-date information to users. The proposed system is a valuable tool for individuals, public health professionals, and policymakers to improve health literacy and support evidence-based decision-making on health topics.

However, the project also highlights some difficulties such as the need for dealing with potential biases or inaccuracies in the data, the complexity of NLP, the development of a multi-agent system to control the real-time data collection and updates to the platform, and ensuring the information provided on the platform is accurate, reliable, and in line with public health guidelines and recommendations. Despite these challenges, the proposed system presents a promising solution to improve health literacy and prevent NCDs, improving public health outcomes and reducing the burden of NCDs on individuals and society.

Further steps for this project may include conducting a pilot study to test the effectiveness of the proposed system and gathering feedback from users. This could involve deploying the system in a real-world setting and assessing its performance in terms of data accuracy. Additionally, further research could be conducted to explore how the system could be adapted to target specific populations or health conditions, or to incorporate additional data sources. Another important step would be to evaluate the system's scalability, in order to adapt to different languages and cultures. Additionally, an additional feature could be included, which would allow the system to predict NCDs based on qualitative data such as diet and sedentary lifestyle and quantitative data such as anthropometric information. This could involve developing a predictive model that uses ML techniques to analyze the data collected from social media, combined with additional information about life habits and physical form. The model could use this information to identify patterns and predict a person's risk of developing an NCD. This feature could also allow users to input their habits and genetics information, and receive personalized recommendations for reducing their risk of developing an NCD. This could be a useful feature for individuals to take proactive measures to improve their health and prevent the onset of NCDs. Finally, it would be important to also consider ethical and legal issues related to data collection and privacy, to ensure that the platform is compliant with relevant regulations and guidelines.

Acknowledgements

This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020. The grants of Daniela Oliveira and Rafaela Rosário are supported by the project "A health promotion intervention for vulnerable school - children and families (BeE-school): a cluster-randomized trial" within the Project Scope PTDC/SAU-ENF/2584/2021.

References

[1] World Health Organization. (2022) "Noncommunicable Diseases (NCD)". [Online]. Available at: https://www.who.int/data/gho/data/themes/noncommunicable-diseases. Accessed: 2022-11-02.

[2] Khatib, O. (2004) "Noncommunicable diseases: Risk factors and regional strategies for prevention and care." *Eastern Mediterranean Health Journal* **10**(6): 778-788.

[3] Davagdorj, Khishigsuren, Bae, Jang-Whan, Pham, Van-Huy, Theera-Umpon, Nipon, and Ryu, Keun Ho. (2021) "Explainable Artificial Intelligence Based Framework for Non-Communicable Diseases Prediction." *IEEE Access* **9** 123672-123688. doi: 10.1109/ACCESS.2021.3110336.

[4] NCD Child. (2022) "Awareness, prevention and treatment of Noncommunicable Diseases (ncds) in Youth." Available at: https://www.ncdchild.org/. Accessed: 2022-12-02.

[5] Moorhead, S., Hazlett, D., Harrison, L., Carroll, J., Irwin, A., and Hoving, C. (2013) "A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication." *J Med Internet Res* doi: 10.2196/jmir.1933

[6] Purnat TD, Vacca P, Czerniak C, et al. Infodemic Signal Detection During the COVID-19 Pandemic: Development of a Methodology for Identifying Potential Information Voids in Online Conversations. JMIR Infodemiology. 2021;1(1):e30971. doi:10.2196/30971.
[7] Goiana da Silva, Francisco, João Marecos, and Francisco Miguel de Abreu Duarte. (2022) "Toolkit for Tackling Misinformation on Noncommunicable Disease: Forum for Tackling Misinformation on Health and NCDs." *World Health Organization*.
[8] "Infodemics and Misinformation Negatively Affect People's Health Behaviours, New WHO Review Finds." (2022) *World Health Organization*.
[9] Bin Naeem, Salman, and Maged N. Kamel Boulos. (2021) "COVID-19 Misinformation Online and Health Literacy: A Brief Overview." *International Journal of Environmental Research and Public Health* **18** (15): 8091.
[10] Lohiniva, Anna-Leena, Anastasiya Nurzhynska, Al-hassan Hudi, Bridget Anim, and Da Costa Aboagye. (2022) "Infodemic Management Using Digital Information and Knowledge Cocreation to Address COVID-19 Vaccine Hesitancy: Case Study from Ghana." *JMIR Infodemiology* **2** (2).
[11] DataReportal. (2022). Digital Around the World - DataReportal – Global Digital Insights. [Online]. Available: https://datareportal.com/global-digital-overview. Accessed: 2022-11-05.
[12] Li, H. O., Bailey, A., Huynh, D., and Chan, J. (2020). YouTube as a source of information on COVID-19: a pandemic of misinformation? BMJ Global Health, 5(5), e002604. https://doi.org/10.1136/bmjgh-2020-002604
[13] Swire-Thompson, B., and Lazer, D. (2020). Public Health and Online Misinformation: Challenges and Recommendations. Annual review of public health, 41, 433–451. https://doi.org/10.1146/annurev-publhealth-040119-094127
[14] CovidCheck. [Online]. Available: https://covidcheck.pt/. Accessed: 2022-11-30.
[15] Klein, A. Z., Magge, A., O'Connor, K., Flores Amaro, J. I., Weissenbacher, D., and Gonzalez Hernandez, G. (2021). Toward Using Twitter for Tracking COVID-19: A Natural Language Processing Pipeline and Exploratory Data Set. J Med Internet Res, 23(1), e25314. https://doi.org/10.2196/25314
[16] WHO. (n.d.). Early AI-supported Response with Social Listening (COVID-19 related conversations online in 30 pilot countries).[Online]. Available: https://www.who-ears.com/. Accessed: 2022-11-05.
[17] Schmaelzle, R., and Wilcox, S. (2022). Harnessing Artificial Intelligence for Health Message Generation: The Folic Acid Message Engine. J Med Internet Res., 24(3), e28858. https://doi.org/10.2196/28858
[18] Purushothaman, V., McMann, T., Nali, M., Li, Z., Cuomo, R., and Mackey, T. K. (2022). Content Analysis of Nicotine Poisoning (Nic Sick) Videos on TikTok: Retrospective Observational Infodemiology Study. J Med Internet Res, 24(3), e34050. https://doi.org/10.2196/34050