Public Health Monitoring System In COVID-19 Conditions Using Machine Learning-Based Sentimental Analysis

Dr. Abhijeet Madhukar Haval[1], Md Afzal[2]

[1]*Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India.*

[2]*Research Scholar, Department of CS & IT, Kalinga University, Raipur, India*

KEYWORDS

Public Health, COVID-19, Machine Learning, Sentimental Analysis

ABSTRACT

Twitter is a significant forum for individuals to discuss and disseminate health-related data. The system offers substantial data for immediate monitoring of contagious diseases (such as COVID-19), relieving disease-prevention organizations from the laborious tasks associated with Personal Health Measures (PHM). PHM identification is a crucial technique for staying informed about the status of an epidemic. It aims to determine an individual's health by analyzing web text data. This research investigates the process of identifying PHM related to COVID-19 using data from Twitter. The research has constructed a COVID-19 PHM dataset with tweets labeled with four distinct categories of health disorders connected to COVID-19: self-mention (SM), other-mention (OM), awareness, and non-healthcare (NHC). The research achieved favorable outcomes in the PHM identification task. The categorizing results enable prompt health tracking and oversight for digital epidemiology. The study assesses the impact of the attention strategy and training methodology on the predictive capabilities.

1. Introduction

The COVID-19 pandemic has resulted in a significant loss of human life on a global scale and poses exceptional difficulties to global health, agricultural systems, and the workforce [1]. This epidemic is causing substantial economic and societal upheaval. A significant number of individuals are at risk of experiencing undernourishment, with the potential for this figure to increase by as much as 134 million by the conclusion of 2020 [8]. The dynamics of COVID-19, such as mortality rates, contagion variables, duration of virus presence in a country, and first deaths, were elucidated by examining the contrasting responses of social media and stock markets to the aftermath of the widespread virus outbreak [2].

The world has undergone a profound alteration, and the research is now in a perpetual transition state. The way of life, relationships, and communication with others have been permanently altered. The risk of viruses plays a crucial role in transmitting, disseminating, and manipulating information in society [3]. The coronavirus has posed a significant pandemic threat and a wide-ranging challenge in management, preparedness, response, and improvement for governments, health organizations, stakeholders, and media outlets.

The current COVID-19 dilemma creates a unique and unprecedented predicament for healthcare communities [17]. Research has achieved significant progress regarding relocation flexibility, unrestricted international travel, and the advancement and application of Information and Communication Technologies (ICTs) [15]. The research must emphasize the growing interconnectedness of global economies, as seen by the financial advancements and individual expertise.

This paper presents the following contributions. The research constructed the inaugural COVID-19 Twitter corpus for a Public Health Measure (PHM) detection study. The research gathered and provided detailed explanations for over 11,000 tweets, including the four categories of health references. The research formulated the detection of PHM as a problem of text classifications. It used a dual Convolutional Neural Network (CNN) to categorize each tweet into one of the four categories. The attention system effectively exploited the keywords in tweets, overcoming the difficulty posed by informal, colloquial, and ambiguous language in tweets. The proposed strategy maximizes the utilization of data acquired at various training phases and has outperformed conventional model training in outcomes.

Related Works

Sentiment Analysis has been conducted to monitor health and gather information on epidemics [5]. The study process can be broadly categorized as knowledge-driven and data-driven [4]. Due to the high cost and time required to create and annotate a PHM database, knowledge-driven techniques were initially used to explore health surveillance using social media data.

Knowledge-driven techniques utilize a knowledge base or ontologies to deliver pertinent health and medical data. A knowledge base generally comprises the associations between healthcare groups, such as the causal link between diabetes and heart disease and the symptomatic relationship between diarrhea and COVID-19. It encompasses medical regulations derived from practical knowledge to anticipate health occurrences. Using data extraction methods, Fu et al. constructed a medical knowledge base called Bio-Caster [6]. The health-related news was extracted and categorized into various health categories. Named entity identification and semantics role, labeling approaches extract the connection among elements in Bio-Caster. The subsequent procedures involved utilizing a Bio-Caster-derived system to predict forthcoming PHM by surveilling articles utilizing pertinent keywords. Fan et al. utilized a medical database to forecast the presence of relevant elements in a tweet [7]. They employed the associative connection between medical terms to link the discovered entities to the classification characteristics, enabling the identification of new illness-related tweets.

Luo et al. introduced the Hidden Flu-State Tweet Modeling (HF-STM) to identify instances of flu mentioned in tweets [16]. The algorithm utilizes tweet-level symptom characteristics to generate health-related themes from the text database. It can categorize an individual's health status as 'healthy,' 'subjected,' or 'infected' by analyzing the tweets they have sent. Eligüzel et al. utilized decision trees, K-Nearest Neighbour (KNN), and Multi-Layer Perception (MLP) classifications to train models to identify whether a tweet includes PHM experiences [9].

To train, the classification utilizes several features, including sentiment keywords and user states. Alkouz et al. analyzed tweets to identify instances of influenza-like symptoms [10]. Given that health references in tweets are typically expressed in simple language, the researchers established connections between these layman’s statements and the technical terminology associated with influenza used in European nations.

Ravi et al. employed pre-existing word embeddings for encoding tweets [18]. A long-term memory (LSTM) network utilizes embeddings as inputs and outputs to determine whether the tweets contain health concerns. García-Díaz et al. employed word embeddings to represent tweet phrases [12]. They inputted them into a bidirectional LSTM (Bi-LSTM) network to categorize influenza-associated tweets. Brglez et al. introduced a methodology to ascertain whether disease-related terms were employed in a metaphorical sense [13]. This approach involves feeding the output into a Convolutional Neural Network (CNN) to determine whether the tweets contain PHMs. Pavlova et al. utilized feelings to identify metaphorical health references [14]. Word embeddings are used to encode the phrases in a tweet. These encrypted words are inputted into a Bi-LSTM and an emotional recognition module.

In summary, investigating PHM from tweets has garnered significant interest. No scientific investigation has been conducted to identify COVID-19 using PHM systems [11]. This is mainly because there is a need for more annotated datasets available for the four universally recognized categories of PHMs. COVID-19 PHMs exhibit a greater degree of imbalance than other diseases, primarily because of the overwhelming number of tweets falling within the 'awareness' group.

Proposed Machine Learning-based PHM

The research utilizes a Machine Learning (ML) methodology to determine the correlation between the tweet and label areas. This approach is chosen for its exceptional ability to extract semantic text data. Complex pre-trained language frameworks have performed excellently in several natural language processing applications. They demand significant computational resources and require substantial training information. The size of the annotation tweets dataset is moderate, with a total of 11,231 tweets. The dimension of the tweets could be more extensive, which suggests that the complex language modeling needs to perform better. This research intends to utilize CNN designs, which have demonstrated exceptional efficacy in extracting regional and universal contextual data from textual data.

The splitting of tweets across the four groups exhibits a significant imbalance, with a minority of Twitter users posting COVID-19-related healthcare remarks regarding themselves and their acquaintances. ML researchers have recognized that imbalanced classes substantially impact the dependability and excellence of ML task outcomes. A classification trained on an imbalanced dataset prioritizes classes with excessive training specimens. The architecture of the proposed research is shown in Fig. 1.

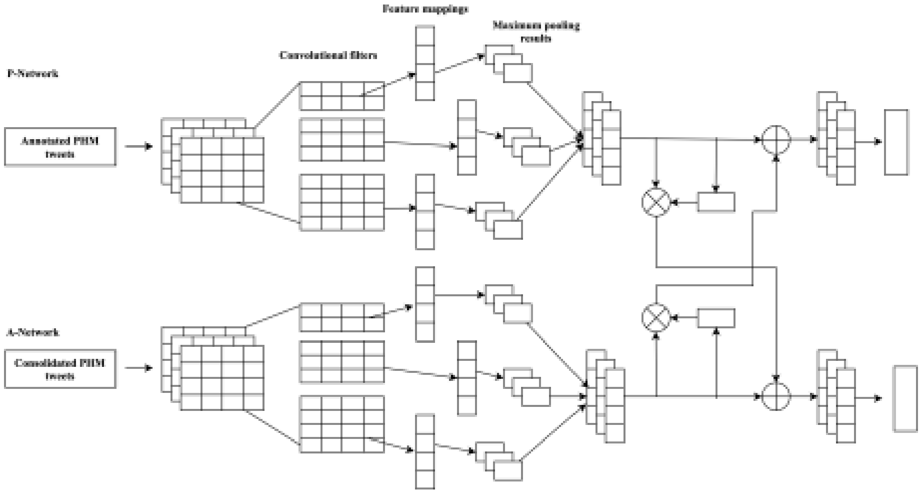

Fig. 1. Architecture of the proposed work

To address this problem, the research suggests implementing a dual CNN architecture. This architecture includes a primary system called Primary Network (P-Network), which categorizes tweets into four specific PHM categories. Given that awareness-class PHM specimens significantly impact the overall PHM dataset, the research wants to enhance the annotated PHM dataset by merging the categories of self-mention (SM), other-mention (OM), awareness, and non-healthcare (NHC). This will help achieve a more balanced distribution of sample sizes throughout the categories. An Auxiliary Network (A-Network) is trained to utilize the aggregated information set to categorize tweets into two categories: the majority, the awareness, and the remainder. Using a more balanced database reduces the likelihood of classifying specimens in the CM category into the awareness category. The result of the A-Network is connected to the P-Network to enhance the understanding of the data distribution across multiple categories.

The P-Network and the A-Network share the same network topology in this research. They utilize distinct training sets and carry out separate categorization tasks. The A-Network could mitigate the categorization bias that arises from unbalanced classes. The purpose of the A-Network is to equalize the P-Network so that it does not favor the dominant category.

Simulation Analysis and Outcomes

The research constructed a corpus of tweets related to COVID-19 to identify information on PHM. The corpus consists of 11,231 tweets published between January and May 2024. The tweets were gathered using keywords such as 'COVID,' 'SARS,' 'coronavirus,' 'coronavirus,' 'pandemic,' and 'quarantines.' The research improved the tweets by eliminating the mentions, hashtags, and links and retaining only the pertinent textual information—two annotators with relevant medical and PHM expertise conducted separate annotations on the 11,231 tweets. The research employed Fleiss's kappa to assess the level of consensus among the annotators regarding inter-annotator consensus. Fleiss's kappa coefficient was calculated to be 0.76, indicating a significant level of consensus in the annotations. The tweets labeled with different categories underwent additional conversations with two annotation experts and one consolidator with experience in PHM to address any annotation disagreements.

The tweets are encoded using word embeddings, which consist of 300 dimensions for each word. Throughout the training phase, the entire set of tweets is divided randomly into three parts: the training set, the verification set, and the assessment set. These parts are allocated in the ratio of 8:1:1, accordingly.

Table 1. Performance analysis

Model	**Label**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F score (%)**
LSTM	SM	53.003	54.044	47.871	35.473
	OM	57.448	62.78	41.601	48.211
	Awareness	72.955	58.829	52.99	52.706
	NHC	45.71	52.743	56.871	46.394
Proposed	SM	79.163	72.411	64.568	74.103
	OM	84.149	83.305	72.348	82.275
	Awareness	85.724	79.014	84.485	79.606
	NHC	63.274	81.268	76.118	85.629

Table 1 demonstrates that class unbalance significantly impacts the detection of PHMs. An immediate outcome is that the result is contingent upon the prevailing class (in this investigation, awareness) and is unrelated to PHM. The classification that has been trained shows a preference for the dominant PHM category and needs to achieve better performance for the pertinent PHM categories (SM and OM). The limited sample sizes of the relevant categories hinder their ability to impact the overall results significantly. The proposed CNN can partially address this problem, as discussed in the previous subsection.

Table 2. Performance analysis of single-layer and double-layer CNN

Model	**Label**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F score (%)**
CNN	SM	43.459	36.737	45.783	38.661
	OM	38.236	42.498	42.934	42.317
	Awareness	42.972	47.954	46.12	47.927
	NHC	36.061	53.822	38.434	52.764
Proposed	SM	54.729	58.098	64.855	75.244
	OM	63.869	74.667	74.508	76.63
	Awareness	75.219	79.11	85.394	83.347
	NHC	89.449	82.259	87.3	86.317

Upon examining Table 2, the proposed model surpasses the CNN in accurately classifying SM, OM, and NHC, as indicated by the total F score for different sets of twitter samples. Although other PHM assignment focuses exist, CNN performs somewhat superior in detecting the awareness category. A comparison of the recollections of these four PHM categories shows that the detection efficiency of the dominant category could be better. The results of the categories with fewer samples show significant improvement. The A-Network offers a proficient method for recognizing PHMs by ensuring equitable efficiency across every category. This is advantageous when the relevant categories have limited samples, as in the PHM detection problem.

2. Conclusion and future scope

This study aims to automate the detection of PHM related to COVID-19 using sophisticated ML and natural language processing methods. The research compiled a corpus of COVID-19 tweets consisting of 11,231 tweets that have been annotated. Every tweet was categorized based on NHC, awareness, SM, and OM criteria. The discovery of COVID-19 PHM was approached as a work of text categorization. A tweet classification model was developed using a CNN approach to categorize each tweet into one of four classifications. Significant progress has been made in terms of the aggregate F score.

Additional studies have been undertaken to examine the impact of the amount of training information on results. The approaches exhibited a bias towards categories with bigger training specimens. Categories with more significant information had increased dependability and improved classification results. The research has demonstrated the suggested technique's strong generalization capabilities through rigorous experimentation. The model proposed utilizing the previous data set can be used on the new social media dataset with acceptable results.

Reference

[1] Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W. C., Wang, C. B., & Bernardini, S. (2020). The COVID-19 pandemic. Critical reviews in clinical laboratory sciences, 57(6), 365-388.

[2] Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., ... & Scala, A. (2020). The COVID-19 social media infodemic. Scientific reports, 10(1), 1-10.

[3] Gulati, A., Pomeranz, C., Qamar, Z., Thomas, S., Frisch, D., George, G., ... & Sundaram, B. (2020). A comprehensive review of manifestations of novel coronaviruses in the context of deadly COVID-19 global pandemic. The American journal of the medical sciences, 360(1), 5-34.

[4] S. Neelima, Manoj Govindaraj, Dr.K. Subramani, Ahmed ALkhayyat, & Dr. Chippy Mohan. (2024). Factors Influencing Data Utilization and Performance of Health Management Information Systems: A Case Study. Indian Journal of Information Sources and Services, 14(2), 146–152. https://doi.org/10.51983/ijiss-2024.14.2.21

[5] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.

[6] Aljeri, Noura, and Azzedine Boukerche. "Mobility management in 5G-enabled vehicular networks: Models, protocols, and classification." ACM Computing Surveys (CSUR) 53.5 (2020): 1-35.

[7] Fu, Z., Zhang, M., Meng, Z., Shen, Y., Buckeridge, D., & Collier, N. (2024, March). BAND: Biomedical Alert News Dataset. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 16, pp. 18012-18020).

[8] Fan, B., Fan, W., & Smith, C. (2020). Adverse drug event detection and extraction from open data: A deep learning approach. Information Processing & Management, 57(1), 102131.

[9] Amiruzzaman, M., Islam, M. R., Islam, M. R., & Nor, R. M. (2022). Analysis of COVID-19: An infectious disease spread. Journal of Internet Services and Information Security, 12(3), 1-15.

[10] Eligüzel, N., Çetinkaya, C., & Dereli, T. (2020). Comparison of different machine learning techniques on location extraction by utilizing geo-tagged tweets: A case study. Advanced Engineering Informatics, 46, 101151.

[11] Alkouz, B., Al Aghbari, Z., Al-Garadi, M. A., & Sarker, A. (2022). Deepluenza: Deep learning for influenza detection from Twitter. Expert Systems with Applications, 198, 116845.

[12] Ahmad, A.S., Ahed, A., Al-smadi, M.K., & Al-smadi, A.M. (2024). Smart Medical Application of Deep Learning (MUNet) for Detection of COVID-19 from Chest Images. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 15(1), 133-153.

[13] García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R., & Valencia-García, R. (2021). Detecting misogyny in

Spanish tweets. An approach based on linguistics features and word embeddings. Future Generation Computer Systems, 114, 506-518.

[14] Brglez, M., Zayed, O., & Buitelaar, P. (2024). TCMeta: a multilingual dataset of COVID tweets for relation-level metaphor analysis. Language Resources and Evaluation, 1-39.

[15] Pavlova, A., & Berkers, P. (2022). “Mental health” defined by Twitter: Frames, emotions, stigma. Health communication, 37(5), 637-647.

[16] Mishna, F., Milne, E., Bogo, M., & Pereira, L. F. (2021). Responding to COVID-19: New trends in social workers’ information and communication technology use. Clinical Social Work Journal, 49, 484-494.

[17] Luo, L., Wang, Y., & Liu, H. (2022). COVID-19 personal health mention detection from tweets using a dual convolutional neural network. Expert Systems with Applications, 200, 117139.

[18] Kutlu, Y., & Camgözlü, Y. (2021). Detection of coronavirus disease (COVID-19) from X-ray images using deep convolutional neural networks. Natural and Engineering Sciences, 6(1), 60-74.

[19] Ravi, J., & Kulkarni, S. (2023). Text embedding techniques for efficient clustering of Twitter data. Evolutionary Intelligence, 16(5), 1667-1677.