Predicting Cardiovascular Disease Risk Using Web-Text Sentiment Analysis and Hybrid Deep Learning Models

Abhijeet Madhukar Haval [1]*, Md Afzal [1]

Abstract

Background: Sentiment Analysis (SA) has emerged as a key tool within Natural Language Processing (NLP) for quantifying emotions expressed in Web Text (WT). Its application in healthcare, particularly in predicting cardiovascular diseases (CD), shows promise. Sentiments related to stress, anger, or other emotional states have been linked to increased CD risk, making Web-Text Sentiment Analysis (WT-SA) a valuable tool in public health prediction. This study compares WT-SA to demographic data from the Centers for Disease Control and Prevention (CDC) in predicting CD risks. Methods: Twitter data was analyzed using SA techniques to assess sentiments related to CD. A hybrid deep learning (DL) model combining 3D-Convolutional Neural Networks (3D-CNN) and Recurrent Neural Networks (RNN) was utilized alongside traditional machine learning (ML) models: Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), and CatBoost. The WT data, spanning three years (2020-2023), was split into training (75%) and testing (25%) datasets. Model performance was evaluated using Accuracy, Precision, Recall, and F1 score. Results: The hybrid 3D-CNN+RNN model achieved the highest test accuracy of 0.95 on Twitter data, outperforming all other models. SVM, LR, NB, and CatBoost achieved accuracies of 0.89, 0.88, 0.75, and 0.77, respectively. When applied to the CDC dataset, the hybrid model reached an accuracy of 0.678, still outperforming the other models. Overall, the WT-SA model demonstrated superior performance compared to using demographic data alone. Conclusion: WT-SA, when combined with hybrid DL models, is a more effective method for predicting CD risk than demographic-based models. This study determined the potential of NLP and DL techniques in leveraging social media data for public health monitoring, suggesting that WT-SA could be a valuable tool for predicting CD risks and enhancing early intervention efforts.

Keywords: Sentiment Analysis, Cardiovascular diseases, Web-Text, Convolution Neural Network, Recurrent Neural Network, Natural Language Processing.

Significance | This study demonstrated the superior potential of Web-Text Sentiment Analysis in predicting cardiovascular disease over demographic-based models.

*Correspondence. Abhijeet Madhukar Haval
Department of CS & IT, Kalinga University,
Raipur, India.
E-mail: md.afzal@kalingauniversity.ac.in

Editor Surendar Aravindhan, Ph.D., And accepted by the Editorial Board Sep 08, 2024 (received for review Jul 05, 2024)

Introduction

Sentiment Analysis (SA), also known as "opinion harvesting," is an emerging field in Natural Language Processing (NLP) that evaluates emotions and thoughts in written texts. SA tools categorize information into three emotional categories: positive, negative, and neutral, while also offering the potential for deeper analysis of written language (Ligthart et al., 2021). Various computational approaches, including artificial intelligence (AI) and machine learning (ML), have been introduced to enhance the accuracy of diagnoses and treatment effectiveness (Briganti & Le Moine, 2020).

Author Affiliation.
[1] Department of CS & IT, Kalinga University, Raipur, India.

Please cite this article.
Abhijeet Madhukar Haval, Md Afzal (2024). Predicting Cardiovascular Disease Risk Using Web-Text Sentiment Analysis and Hybrid Deep Learning Models, Journal of Angiotherapy, 8(9), 1-5, 9876

https://doi.org/10.25163/angiotherapy.899876

However, because intuition is vital in patient care, advancements in medical AI/ML must incorporate computational techniques that detect and evaluate emotions to address health issues more effectively. A study utilizing SA analyzed physicians' written notes in critical care, revealing that doctors' intuitive judgments, or "gut feelings," were pivotal in shaping patient treatment outcomes (Brezulianu et al., 2022).

A recent review evaluated the effectiveness of SA techniques in analyzing healthcare-related emotions on Twitter, but no study has specifically examined its application in cardiology (Gohil et al., 2018; Lavanya et al., 2024). Cardiovascular diseases (CD) remain a global concern, being among the leading causes of death worldwide (Laranjo et al., 2024). Emerging evidence suggests that SA could prove valuable in cardiology, especially in the growing field of telemedicine (Eberly et al., 2020).

Social media's increasing popularity offers a rich foundation for developing SA models applicable to cardiology. For instance, a study on medication safety showed that SA enhanced state-of-the-art methods in detecting adverse drug reactions (ADR), improving the F1-score for identifying hostile responses in a Twitter dataset related to 82 medications, including cardiovascular treatments (Fan et al., 2020). The relationship between CD and mental health disorders is reciprocal, and emotional disruptions can exacerbate coronary artery disease, making emotional state evaluation essential in CD risk assessment (Toshtemirovna et al., 2022). However, the subjective nature of emotions and technological limitations have hindered the incorporation of emotional state assessments in predicting CD risks (Gümüş et al., 2022). Technological advancements in AI and SA present an opportunity to systematically evaluate emotions, providing a more accurate understanding of patients' health through online platforms (Yazdani et al., 2023).

The rise of social media and forums facilitates the collection of vast amounts of unstructured text, which can be used to experiment with innovative SA techniques. One study analyzed 5 million tweets about cardiovascular diseases, covering themes such as risk factors and treatment (Huang et al., 2020). This research utilized Twitter data to predict heart disease mortality at the state level by analyzing language patterns. SA outperformed traditional predictive models, leading to more accurate and complex CD forecasts (Wankhade et al., 2022).

A key challenge in advancing ML models is that 91% of global information is unstructured, making it difficult to apply conventional AI/ML techniques (Mahadevkar et al., 2024). While structured data like test results are well-organized, observations, intuitions, and experiences remain disordered, often recorded in medical archives or online platforms. SA's ability to systematically analyze unstructured data positions it to overcome these challenges, offering improved outcomes for understanding patient experiences.

2. Materials and Methods

Natural language processing (NLP) techniques, including sentiment analysis (SA), are capable of extracting complex sentiments such as ratings and classifiers, which provide notable benefits for cardiology research. Emotions, as demonstrated in several studies, are significant risk factors in the development of cardiovascular diseases (CD) (Mumtaj Begum, 2022). Researchers have employed Twitter sentiment analysis (WT-SA) to predict the likelihood of CD by analyzing thoughts shared on social media platforms like Twitter. A new vocabulary related to CD has been developed through the analysis of sentiments expressed in tweets, with machine learning (ML) algorithms categorizing individuals' CD risks. These algorithms were applied to a database from the Centers for Disease Control and Prevention (CDC), which included demographic data for comparison purposes. One study proposed using a 3D convolutional neural network (CNN) in combination with a recurrent neural network (RNN) based on hybrid deep learning (DL) models to forecast CD using WT-SA. Figure 1 illustrates the framework for CD prediction based on WT-SA (Sathyanarayanan & Srikanta, 2024).

2.1 Keywords Selection and Twitter Data Extraction

The use of language analysis to comprehend psychological states has a long history, with conventional methods relying on dictionary lists containing terms associated with specific emotions. For example, a negative-emotion dictionary might include words such as "sad," "gloomy," and "weeping." In this study, psychiatric and clinical terms related to CD found in tweets were utilized to identify individuals at risk of developing CD. Keywords included common medical terms such as "anaesthesia," "angiography," "myocardial infarction," "hypertension," and "tachycardia." Recognized CD risk factors like smoking, high cholesterol, and stress were also included. According to the CDC, excessive alcohol use is another critical risk factor for CD. Though more keywords could be added, this study was limited to a set of 10 terms. These keywords were then employed to extract data from Twitter, which was analyzed and inputted into an algorithm designed to detect or predict potential CD risks.

To collect data, the study utilized Tweepy, an API for gathering information from Twitter (Tweepy, 2024). After obtaining the necessary credentials, tweets were extracted based on specific keywords, locations, and time frames. Over 301,845 tweets from ten U.S. states were gathered between 2020 and 2023. While collecting Twitter data based on selected keywords is common practice, this study distinguished itself by curating a precise set of terms related to CD. Unlike conventional research, the chosen keywords captured the nuanced linguistic patterns of Twitter users discussing

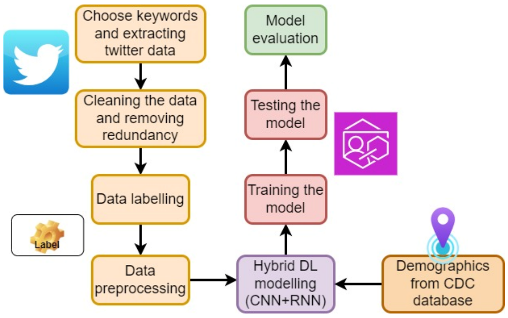
Choose keywords and extracting twitter data
Cleaning the data and removing redundancy
Data labelling
Label
Data preprocessing
Hybrid DL modelling (CNN+RNN)
Demographics from CDC database
Training the model
Testing the model
Model evaluation

Figure 1 Framework of the proposed CD prediction using WT-SA

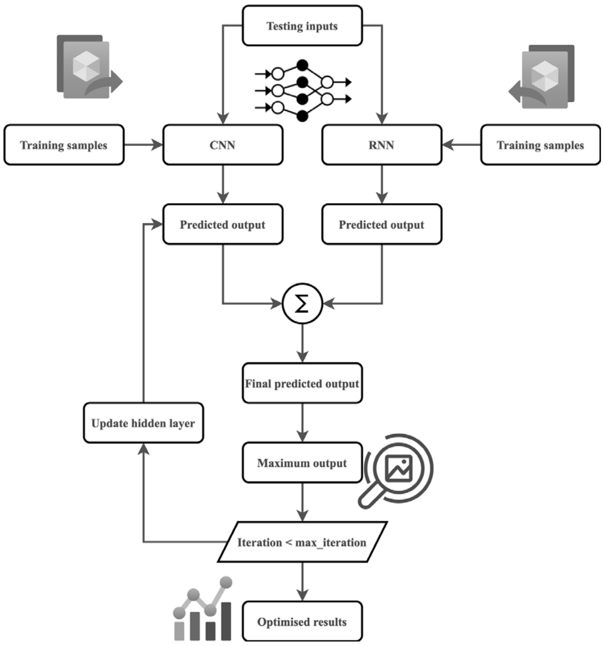

Figure 2. Hybrid DL (3D-CNN+RNN) framework

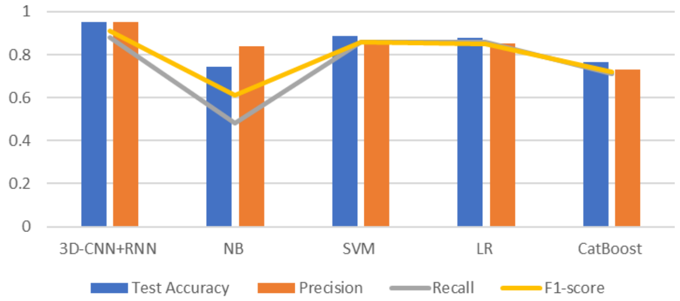

Figure 3(a) Performance analysis of CD prediction using WT-SA for various ML/DL algorithms

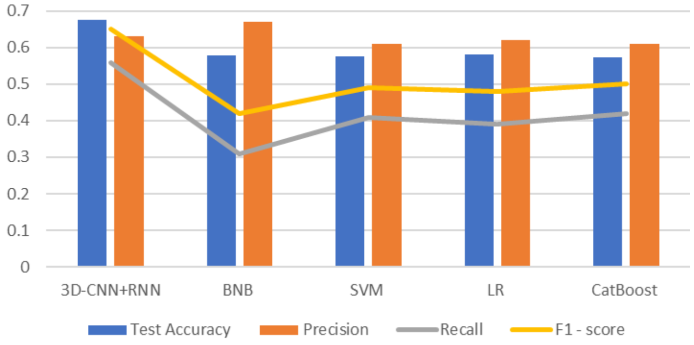

Figure 3(b) Performance analysis of CD prediction using CDC dataset for various ML/DL algorithms

https://doi.org/10.25163/angiotherapy.899876

their experiences with CD. The dictionary included official clinical terminology and everyday language, capturing a broader, more authentic depiction of CD discussions on social media.

2.2 Data Preprocessing

Data preprocessing in natural language processing (NLP) is a crucial step that involves cleaning and organizing data for deep learning (DL) applications. This process typically includes tokenization, converting text to lowercase, removing word endings, and sorting. In this study, the VADER algorithm was used to analyze emotions expressed in tweets to assess the potential risk of cardiovascular disease (CD) (Elbagir & Yang, 2019). A cutoff value of -0.25 was set to detect favorable sentiments related to CD risk. Users whose sentiment scores exceeded this threshold were labeled '1', indicating a possible CD risk, while those below the threshold were labeled '0', indicating no risk. Sentiment analysis (SA) was employed to evaluate the model's performance by categorizing users into these classes. The database was then divided into a 75%-25% ratio for training and testing.

The decision to use a threshold of -0.25 in the VADER model was based on initial experimentation with a small dataset, where the cutoff value closely aligned with physical classification procedures. This threshold was subsequently used to categorize a larger training dataset, serving as the foundation for comparing different DL models' predictive capabilities regarding CD risk.

2.3 Hybrid DL Modeling (3D-CNN + RNN)

Twitter data was used to predict and classify users into two categories: healthy and at-risk for CD, using a hybrid approach combining 3D convolutional neural networks (CNN) and recurrent neural networks (RNN). The proposed model implemented this hybrid CNN + RNN framework with an enhanced hidden layer (HL) technique. As shown in Figure 2, the hybrid DL-based CD detection framework first extracted inputs from Twitter data, which were then preprocessed. CNN and RNN models were employed for learning trials, and the test results were evaluated using the hybrid approach. The outputs from both CNN and RNN were fused to produce a final prediction. This process was iterated until optimal outcomes were achieved, with feedback incorporated into the HL weights for improvement.

The model construction involved six convolution layers (CL), four interconnecting layers, four max-pooling layers, and two fully connected layers. The first CL applied 96 filters of size 3×3×3 to the input data. Max-pooling, with filters of 2×2×2, reduced the output size after each CL. The remaining layers processed the downsampled feature maps to refine the predictions. Fully connected layers linked the neurons from the previous levels, while RNNs processed input data through each stage by reusing output from the previous phase. Applying the 3D-CNN+RNN model for CD detection proved to be a valuable approach for early recognition, treatment, and management of cardiovascular diseases.

3 Results and Discussion

Demographic data, including ethnic background and gender, were numerically encoded using one-hot encoding for the three-year period between 2020 and 2023. The focus of the analysis was to predict individuals at risk of cardiovascular disease (CD). Five models were utilized: four traditional machine learning (ML) models—Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), and CatBoost—and one hybrid deep learning (DL) model combining 3D convolutional neural networks (3D-CNN) and recurrent neural networks (RNN). The dataset was split into a 75%-25% ratio for training and testing.

Figure 3(a) illustrates the performance outcomes of the sentiment analysis of Twitter data (WT-SA), which involved tweets collected from ten states over three years. The hybrid 3D-CNN+RNN model achieved the highest accuracy of 0.95. The SVM followed with an accuracy of 0.89, and LR achieved 0.88. NB and CatBoost obtained accuracies of 0.75 and 0.77, respectively. Figure 3(b) shows the results for the CDC dataset, with the hybrid 3D-CNN+RNN model again leading with a test accuracy of 0.678, followed by LR at 0.584, NB at 0.571, SVM at 0.575, and CatBoost at 0.574.

A comparative analysis between the WT-SA and CDC datasets demonstrated that the proposed WT-SA model outperformed the CDC dataset. The study's goal was to evaluate the effectiveness of natural language processing (NLP) and DL techniques on Twitter data to identify individuals at risk of developing CD. The results showed that this method produced better outcomes than relying solely on demographic data.

The study proposed a 3D-CNN+RNN hybrid model based on DL for predicting CD using WT-SA. Performance metrics such as accuracy, precision, recall, and F1 score were assessed, with the hybrid model achieving the best test accuracy of 0.95. SVM followed with an accuracy of 0.89, and LR came close with 0.88. NB and CatBoost achieved test accuracies of 0.75 and 0.77, respectively. The results highlighted that using NLP and ML to analyze WT data can effectively predict individuals at risk of CD, with sentiment analysis (SA) used for label creation. The findings were validated against a CDC demographic dataset to confirm the approach's reliability. Further parameter adjustments can enhance model performance.

The study also noted that, at the state level, WT data predictions outperformed those based on demographic information. This suggests that WT-SA is a superior method for predicting or categorizing individuals at risk for developing CD, compared to demographic-based models.

4 Conclusion

This study demonstrated the effectiveness of Web-Text Sentiment Analysis (WT-SA) in predicting cardiovascular disease (CD) risks, surpassing traditional demographic-based models. By employing a hybrid 3D Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), the model achieved an impressive 95% accuracy in identifying individuals at risk. The analysis of sentiments from Twitter data provided a more nuanced understanding of CD risk factors, revealing that emotional states, captured through sentiment, significantly contribute to CD prediction. The study's results affirm the value of Natural Language Processing (NLP) and machine learning (ML) techniques in healthcare, particularly in cardiology, where emotional factors play a critical role in disease development. In comparison to a demographic dataset from the Centers for Disease Control and Prevention (CDC), the WT-SA model delivered superior performance, suggesting that sentiment analysis can be a valuable tool in early disease detection and health monitoring, providing new avenues for public health initiatives.

Author contributions

AMH and MA contributed to conceptualization, fieldwork, data analysis, drafting the original manuscript, editing, funding acquisition, and manuscript review. Both AMH and MA were also involved in research design, methodology validation, data analysis, visualization, and manuscript review and editing. Additionally, AMH took lead in methodology validation, investigation, funding acquisition, supervision, and final revisions. All authors have reviewed and approved the final version of the manuscript.

Acknowledgment

The authors were thankful to their department.

Competing financial interests

The authors have no conflict of interest.

References

Brezulianu, A., Burlacu, A., Popa, I. V., Arif, M., & Geman, O. (2022). "Not by our feeling, but by others' seeing": Sentiment analysis technique in cardiology—An exploratory review. Frontiers in Public Health, 10, 880207.

Briganti, G., & Le Moine, O. (2020). Artificial intelligence in medicine: Today and tomorrow. Frontiers in Medicine, 7, 509744.

Eberly, L. A., Khatana, S. A. M., Nathan, A. S., Snider, C., Julien, H. M., Deleener, M. E., & Adusumalli, S. (2020). Telemedicine outpatient cardiovascular care during the COVID-19 pandemic: Bridging or opening the digital divide?. Circulation, 142(5), 510-512.

Elbagir, S., & Yang, J. (2019, March). Twitter sentiment analysis using natural language toolkit and VADER sentiment. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 122, No. 16). sn.

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.

Gohil, S., Vuik, S., & Darzi, A. (2018). Sentiment analysis of health care tweets: Review of the methods used. JMIR Public Health and Surveillance, 4(2), e5789.

Gümüş, A. E., Uyulan, Ç., & Guleken, Z. (2022). Detection of EEG patterns for induced fear emotion state via EMOTIV EEG testbench. Natural and Engineering Sciences, 7(2), 148-168.

Huang, D., Huang, Y., Adams, N., Nguyen, T. T., & Nguyen, Q. C. (2020). Twitter-characterized sentiment towards racial/ethnic minorities and cardiovascular disease (CD) outcomes. Journal of Racial and Ethnic Health Disparities, 7(5), 888-900.

Laranjo, L., Lanas, F., Sun, M. C., Chen, D. A., Hynes, L., Imran, T. F., ... & Chow, C. K. (2024). World Heart Federation roadmap for secondary prevention of cardiovascular disease: 2023 update. Global Heart, 19(1).

Lavanya, P., Subba, R. I. V., Selvakumar, V., & Deshpande, S. V. (2024). An intelligent health surveillance system: Predictive modeling of cardiovascular parameters through machine learning algorithms using LoRa communication and the Internet of Medical Things (IoMT). Journal of Internet Services and Information Security, 14(1), 165-179.

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: A tertiary study. Artificial Intelligence Review, 1-57.

Mahadevkar, S. V., Patil, S., Kotecha, K., Soong, L. W., & Choudhury, T. (2024). Exploring AI-driven approaches for unstructured document analysis and future horizons. Journal of Big Data, 11(1), 92.

Mumtaj Begum, H. (2022). Scientometric analysis of the research paper output on artificial intelligence: A study. Indian Journal of Information Sources and Services, 12(1), 52–58.

Reference

Sathyanarayanan, S., & Srikanta, M. K. (2024). Heart sound analysis using SAINet incorporating CNN and transfer learning for detecting heart diseases. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 15(2), 152-169. https://doi.org/10.58346/JOWUA.2024.I2.011

Toshtemirovna, E. M. M., Alisherovna, K. M., Totlibayevich, Y. S., & Xudoyberdiyevich, G. X. (2022). Anxiety disorders and coronary heart disease. The Peerian Journal, 11, 58-63.

Tweepy. (2024). Retrieved from https://docs.tweepy.org/en/stable/

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.

Yazdani, A., Shamloo, M., Khaki, M., & Nahvijou, A. (2023). Use of sentiment analysis for capturing hospitalized cancer patients' experience from free-text comments in the Persian language. BMC Medical Informatics and Decision Making, 23(1), 275.