Cardiovascular Disease Risk Prediction via Social Media

Al Zadid S.B. Habib[1], M.A. Bin Syed[2], M.T. Islam[3], and Donald A. Adjeroh[1]

[1] Lane Department of Computer Science and Electrical Engineering
[2] Department of Industrial and Management Systems Engineering
[3] Wadsworth Department of Civil and Environmental Engineering
West Virginia University, Morgantown WV 26506, USA
{ah00069, ms00110, mi00018}@mix.wvu.edu, donald.adjeroh@mail.wvu.edu

Abstract. Researchers use Twitter and sentiment analysis to predict Cardiovascular Disease (CVD) risk. We developed a new dictionary of CVD-related keywords by analyzing emotions expressed in tweets. Tweets from eighteen US states, including the Appalachian region, were collected. Using the VADER model for sentiment analysis, users were classified as potentially at CVD risk. Machine Learning (ML) models were employed to classify individuals' CVD risk and applied to a CDC dataset with demographic information to make the comparison. Performance evaluation metrics such as Test Accuracy, Precision, Recall, F1 score, Mathew's Correlation Coefficient (MCC), and Cohen's Kappa (CK) score were considered. Results demonstrated that analyzing tweets' emotions surpassed the predictive power of demographic data alone, enabling the identification of individuals at potential risk of developing CVD. This research highlights the potential of Natural Language Processing (NLP) and ML techniques in using tweets to identify individuals with CVD risks, providing an alternative approach to traditional demographic information for public health monitoring.

Keywords: Social Media Analytics · Cardiovascular Disease · Machine Learning · Deep Learning · CDC · CNN-LSTM.

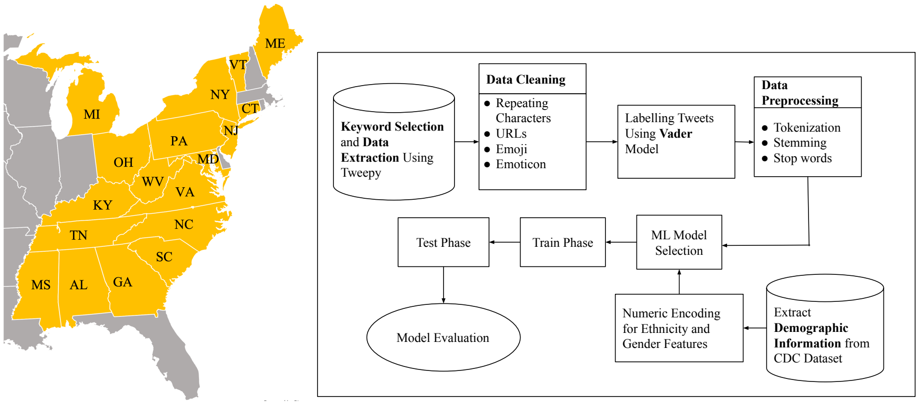
1 Introduction

CVD is one of the most significant health concerns related to morbidity and mortality [18]. For example, heart disease and stroke are the first and third most common fatalities in the US, respectively [16]. A previous study found that one in three Americans have one or more CVDs [20]. Psychological traits have been demonstrated to raise the risk for CVD via physiological consequences and unhealthy behaviors [15]. Social media data now contains a wealth of information about communities' psychological conditions and behaviors. It is a good platform for people to share personal experiences, seek information, and exchange mental support or sympathy on health issues. Thus, through Twitter, researchers can systematically observe public discourse on health issues, such as CVD [25]. In essence, it is achieved through NLP by determining the polarity of word data, which is divided into negative, positive, and neutral categories [10,12,21]. Moreover, existing literature supports the hypothesis that there are no significant differences in the proportion of social media users irrespective of psychological characteristics, which nullifies the concern regarding biased estimates.

Convolutional Neural Network (CNN) is a Deep Learning (DL) architecture that extracts features from the input before sending it through filters. Long Short-Term Memory (LSTM) is a type of neural network that attempts to give the Recurrent Neural Network (RNN) a short-term memory that can endure thousands of timesteps. Thus, the Hybrid CNN-LSTM architecture can classify people into high to low-risk categories. Since NLP-based research is still in its infancy, the risk prediction of CVDs with hybrid CNN-LSTM is yet to be explored. Moreover, several ML algorithms like Bernoulli Naive Bayes (BNB), Support Vector Machine (SVM), Logistic Regression (LR), and CatBoost are known to be good with categorical data.

One of NLP's biggest challenges is finding the appropriate dictionaries for the specific problem. Thus we developed a suitable NLP dictionary for CVD based on current literature. We choose some clinical and psychological risk factors as keywords causing CVD [3], which people might use in their tweets. Then, we employed dictionary-based analysis to define the psychological language correlates of CVDs using hybrid CNN-LSTM architecture and other ML algorithms. We have chosen these keywords for causing CVD, which are considered risk factors to cause CVD or somehow create the possibility to relate to CVD. Based on those keywords, we have collected tweets from eighteen US states, including thirteen Appalachian states [2], namely Alabama (AL), Georgia (GA), Kentucky (KY), Maryland (MD), Mississippi (MS), New York (NY), North Carolina (NC), Ohio (OH), Pennsylvania (PA), South Carolina (SC), Tennessee (TN), Virginia (VA), and West Virginia (WV). Apart from these Appalachian states, five other nearby states, namely Michigan (MI), New Jersey (NJ), Vermont (VT), Maine (ME), and Connecticut (CT), were chosen to extract tweets from those regions based on our selected keywords. Fig. 1(a) depicts the map of the selected states.

In this study, a state-wise CDC dataset from the US Department of Health & Human Services was collected for 2019-2020 [8], containing 124 indicators related to chronic diseases. We applied the same ML models to a CDC dataset containing patients' demographic information in those eighteen states. The dataset contained demographic information like gender, race, ethnicity, location, diagnosis year, and disease type. Considering CVD as a class '1' and the other diseases as a class '0' throws a binary classification problem where we must classify individuals with CVD based on demographic information. The CDC dataset says the number of people who have already developed CVD in those states for that given timeframe. On the other hand, the Twitter-based dataset indicates the number of individuals who might develop CVD or already have CVD in those states for that timeframe. Overall, this paper has focused on Twitter being a great source in such predictions to assist public health practitioners.



(a) Selected States (b) Workflow

Fig. 1. Selected States to Collect Tweets and Overall Workflow.

2 Related Works

Previously, researchers employed NLP to track distinct diseases, i.e., flu, H1N1 influenza, depression, etc. [14, 17, 19, 26]. With the help of multi-source search engine data, Su et al. produced an accurate influenza forecast in regions with erratic seasonal influenza trends [27]. In pilot research, 135,000 tweets were collected during the Swine Flu pandemic in a week to extract communications and detect outbreaks from Twitter data [24]. Besides, Twitter data was utilized for mining public health topics in different studies [22, 23].

In [15], the researchers tried to find a connection between the language used on Twitter and heart disease mortality rates at the county level. By analyzing the psychological language patterns in tweets from over 1,300 U.S. counties, they discover that counties with more negative and less positive language on Twitter tend to have higher heart disease mortality rates. This research suggests that analyzing social media language can provide insights into public health outcomes, enabling the prediction of heart disease mortality and potentially informing targeted interventions (See counterclaim at [13]).

In [25], the researchers tried to investigate the use of Twitter as a data source for CVD research in response to the counterclaim. They analyzed a large sample of tweets to identify CVD-related keywords and hashtags. The study finds a significant correlation between the frequency of these CVD-related terms on Twitter and established risk factors and behaviors associated with cardiovascular health. This suggests that Twitter data has the potential to provide valuable insights into public discussions and behaviors related to CVD. The study highlights the utility of social media platforms like Twitter as a supplementary data source for monitoring and understanding cardiovascular health at a population level, which can inform public health interventions and targeted messaging for promoting cardiovascular well-being.

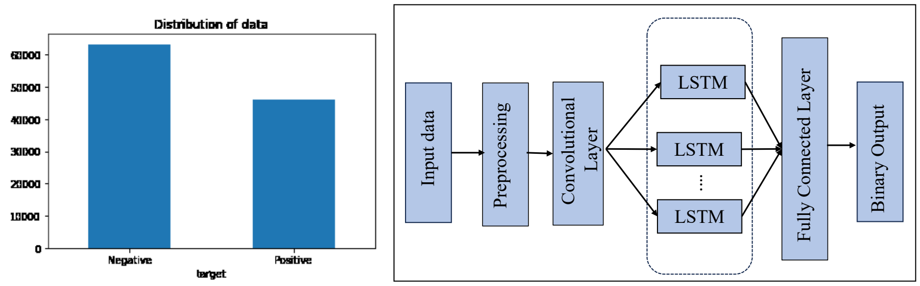
In [15] and [25], the authors tried to find the correlation with CVD regarding demographic information and tweets. In our work, we tried to approach it differently. We developed our dictionary with keywords related to CVD to extract tweets. We used VADER for sentiment analysis to find the polarity. Based on the polarity of the sentiment, we assigned labels to the users with two classes, one with the risks of developing potential CVD and the other not developing any risk of CVD. In contrast to these two papers, we tried to assess this evaluation at the state level comprising eighteen states instead of the county level. Moreover, the timeframe of our collected tweets was also different. We are trying to denote that words expressed on social media can bear different meanings, and based on the risk factors of CVD, we can identify persons who have been developing the CVD.

3 Methodological Framework

Fig. 1(b) can be denoted as the methodological framework for this work as a work flow:

3.1 Tweet-Based Approach

Keywords Selection and Dataset Collection Language analysis techniques for interpreting psychological states have a long history. Dictionary lists of terms connected to various structures are used in conventional methods (e.g., sad, glum, and crying are part of a negative-emotion dictionary). We have used the psychological and clinical keywords used in tweets to identify the persons with potential CVD risk. The keywords we have chosen are based on several categories. Anesthesia, angiogram, cardiologist, echocardiogram, heart attack, heart failure, hypertension, and chest pain are some of the most commonly used cardiac terminologies, according to Columbia Heart Surgery [3]. Smoking and cholesterol are two of the most known risk factors for heart-related diseases as per CDC [4]. Stress is considered to be a key psychological factor for causing CVD [11], and alcohol use in excessive amounts can be a leading reason to cause CVD, according to John Hopkins Medicine [1]. More keywords or terminologies can be selected based on different criteria, but we decided to stick to twelve as a standard number. The keywords are then used for extracting the data from Twitter to analyze and feed into the model for identifying or predicting persons with potential CVD risks. Tweepy [7] is used to collect data from the Twitter API. By obtaining a developer account and the necessary access keys, tweets can be extracted based on specific keywords, locations, and periods. In this study, 269,969 tweets were collected over three years (2019 to 2021) from eighteen US states using twelve keywords. Collecting Twitter data using a specific set of keywords is indeed a common practice. However, the novelty of our study lies precisely in curating these CVD-related keywords. While traditional studies may rely on a broader range of CVD-related terms, our method stands out because it uses a carefully selected set of keywords that capture social media users' complex language patterns when discussing their cardiovascular disease experiences. Our keyword dictionary contains more than just formal clinical terms for CVD. We also included everyday language and lifestyle-related phrases often found in Twitter discussions about CVD. This allows our model to capture a broader and more realistic picture of how people talk about CVD on social media.



(a) Class Distribution

(b) CNN-LSTM Architecture

Fig. 2. Twitter Data Class Distribution and CNN-LSTM Hybrid Architecture.

Dataset Preprocessing and Polarity Identification Dataset preprocessing in NLP is a crucial step involving cleaning and formatting data before using it in ML or DL models. Preprocessing includes tokenizing, lowercasing, removing stop words and punctuation, and stemming. This study uses the VADER model to analyze sentiment in tweets and assess the potential risk of CVD. A threshold of -0.30 is set to identify positive sentiments related to CVD risks. Users above the threshold are labeled as '1' for potential risk, while those below are labeled as '0' for no risk [5,9]. Model performance is evaluated based on sentiment analysis results, and the class distribution indicates sentiment labels ('1' for positive and '0' for negative). Fig. 2(a) shows the class distribution after applying the VADER model. The dataset is split into a 70%-30% ratio for training and testing after labeling based on VADER analysis. In our study, we decided to use the VADER model with a threshold of -0.30 based on our preliminary work with a smaller dataset. We observed that this threshold closely aligned with our manual labeling process. It is important to note that this choice was made empirically, and we applied this threshold to label our larger training dataset. However, our main objective was not to assert that the VADER model with this specific threshold is universally the best option. Instead, we intended to use this labeled dataset as a reference for assessing and comparing the predictive performance of various ML and DL models in predicting CVD risks.

Performance Evaluation with ML and DL This work utilizes a combination of ML and DL models. A hybrid CNN-LSTM model is constructed for the

DL approach, as shown in Fig. 2(b). Conventional ML algorithms such as BNB, SVM, LR, and CatBoost are employed. Labels are assigned based on sentiment polarity identified by VADER, with '1' indicating positive sentiment and potential CVD risks and '0' indicating negative sentiment and no CVD risks. The dataset is split into training and test sets, with the ML and DL models trained on the training set and evaluated on the test set. The aim is to predict individuals with CVD risk based on sentiment expressed on Twitter. Performance evaluation metrics, including test accuracy, precision, recall, F1, MCC, and CK scores, are used to assess the performance of the classifiers.

3.2 CDC Dataset-Based Approach

Demographic information, including gender and ethnicity, was encoded numerically. One-hot encoding was used for the years 2019 and 2020. The focus was on predicting individuals with CVD. The dataset was balanced using the SMOTE technique [6]. Four ML models (BNB, SVM, LR, CatBoost) and a CNN-LSTM hybrid DL model were used. The dataset was split at 70%-30% ratio for training and testing. Limited demographic features were a challenge despite the large number of patients.

4 Results Analysis

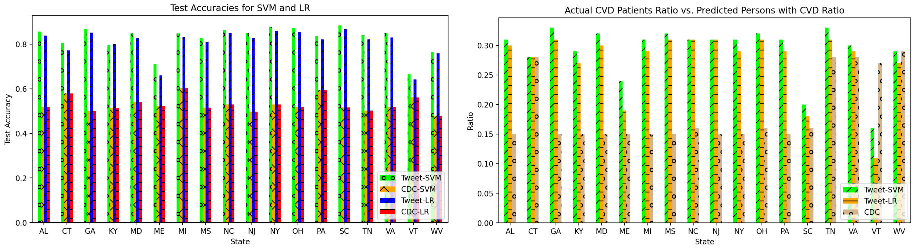
Table 1 shows the results for the overall Twitter dataset, consisting of tweets from eighteen states over three years. The CNN-LSTM hybrid model achieved 77.51% test accuracy. SVM ranked highest with 88.75% accuracy, followed by LR with 87.82%. BNB and CatBoost achieved test accuracies of 74.55% and 76.67%, respectively. Table 2 presents the results for the CDC dataset, where LR ranked highest with 58.03% test accuracy, followed by BNB with 57.93%. CNN-LSTM, SVM, and CatBoost achieved test accuracies of 57.64%, 57.55%, and 57.42%, respectively. Comparative performance evaluations were conducted between the Twitter and CDC datasets, indicating that the Twitter-based dataset generally outperforms the CDC dataset. The study aims to establish the effectiveness of using NLP and ML with Twitter data to identify individuals at risk of CVD, which proves more effective than relying on demographic information alone. Fig. 3(a) illustrates the state-wise comparative performances of SVM and LR for Twitter and CDC datasets regarding test accuracy as the best two classifiers for Twitter dataset. Ratios comparing actual and predicted CVD prevalence based on CDC and Twitter data offer insights into the potential effectiveness of using social media data for predicting and estimating CVD prevalence which is outlined in Fig. 3(b). While these findings suggest the utility of social media data for public health research, further analysis and validation are required to ensure the reliability and generalizability of predictions derived from Twitter data.

Table 1. Performance Evaluation for the Overall Twitter Based Dataset

Model	**Test Accuracy**	**Precision**	**Recall**	**F1**	**MCC**	**CK**
CNN-LSTM	77.51%	0.75	0.68	0.72	0.53	0.53
BNB	74.55%	0.84	0.48	0.61	0.48	0.44
SVM	88.75%	0.87	0.86	0.86	0.77	0.77
LR	87.82%	0.85	0.86	0.85	0.75	0.75
CatBoost	76.67%	0.73	0.71	0.72	0.52	0.53

Table 2. Performance Evaluation for the CDC Dataset

Model	**Test Accuracy**	**Precision**	**Recall**	**F1**	**MCC**	**CK**
CNN-LSTM	57.64%	0.63	0.36	0.45	0.17	0.15
BNB	57.93%	0.67	0.31	0.42	0.19	0.16
SVM	57.55%	0.61	0.41	0.49	0.16	0.15
LR	58.03%	0.62	0.39	0.48	0.17	0.16
CatBoost	57.42%	0.61	0.42	0.50	0.16	0.15



(a) Test Accuracy

(b) Ratio

Fig. 3. Results for SVM and LR.

5 Conclusions

This study collected tweets from eighteen US states using a new dictionary. Before conducting sentiment analysis, preprocessing techniques like stemming and lemmatization were applied to the tweets. The VADER model was then used to determine the polarity of the tweets (negative or positive) and assign labels for ML and DL models. ML models such as BNB, SVM, LR, and CatBoost were utilized, along with the CNN-LSTM hybrid model for DL. The SVM achieved the highest test accuracy of 88.75%, followed by LR with 87.82% accuracy. CNN-LSTM achieved 77.51% test accuracy. The study demonstrates that NLP and ML can effectively use Twitter data to predict individuals with potential CVD risk, with sentiment analysis aiding in label creation. Results were compared with the CDC dataset with demographic information to validate the claim. Additionally, performance evaluation parameters can be incorporated to assess the models further. State-wise performance was comparatively better for tweets than demographic information for the best two classifiers. This indicates that tweets are a valuable resource for predicting or classifying individuals at risk of CVD compared to demographic data.

References

1. Alcohol and heart health: Separating fact from fiction | johns hopkins medicine. https://www.hopkinsmedicine.org/health/wellness-and-prevention/alcohol-and-heart-health-separating-fact-from-fiction, (Accessed on 12/12/2022)
2. Appalachian states - appalachian regional commission. https://www.arc.gov/appalachian-states/, (Accessed on 12/12/2022)
3. Glossary of words related to heart disease: Cardiac surgery program | columbia university department of surgery. https://columbiasurgery.org/heart/glossary, (Accessed on 12/12/2022)
4. Know your risk for heart disease | cdc.gov. https://www.cdc.gov/heartdisease/risk_factors.htm, (Accessed on 12/12/2022)
5. Sentimental analysis using vader. interpretation and classification of... | by aditya beri | towards data science. https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664, (Accessed on 12/12/2022)
6. Smote for imbalanced classification with python - machinelearningmastery.com. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/, (Accessed on 12/13/2022)
7. Tweepy documentation — tweepy 4.12.1 documentation. https://docs.tweepy.org/en/stable/, (Accessed on 12/12/2022)
8. U.s. chronic disease indicators (cdi) - catalog. https://catalog.data.gov/dataset/u-s-chronic-disease-indicators-cdi, (Accessed on 12/13/2022)
9. Vader sentiment analysis | nlp sentiment analysis using vader. https://www.analyticsvidhya.com/blog/2021/06/vader-for-sentiment-analysis, (Accessed on 12/12/2022)
10. Abdel-Baki, A., et al.: Understanding access and use of technology among youth with first-episode psychosis to inform the development of technology-enabled therapeutic interventions. Early Intervention in Psychiatry **11**(1), 72–76 (2017)
11. Albus, C.: Psychological and social factors in coronary heart disease. Annals of medicine **42**(7), 487–494 (2010)
12. Berryman, C., et al.: Social media use and mental health among young adults. Psychiatric quarterly **89**(2), 307–314 (2018)
13. Brown, N.J., Coyne, J.C.: Does twitter language reliably predict heart disease? a commentary on eichstaedt et al.(2015a). PeerJ **6**, e5656 (2018)
14. Chew, C., Eysenbach, G.: Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. PloS one **5**(11), e14118 (2010)
15. Eichstaedt, J.C., et al.: Psychological language on twitter predicts county-level heart disease mortality. Psychological science **26**(2), 159–169 (2015)
16. for Health Statistics, N.C., et al.: Table 31. death rates for motor vehicle-related injuries, by sex, race, hispanic origin, and age: United states, selected years 1950–2013. Health, United States pp. 55–64 (2014)

17. Hussain, J., et al.: Exploring the dominant features of social media for depression detection. Journal of Information Science **46**(6), 739–759 (2020)
18. Kawachi, I., et al.: A prospective study of social networks in relation to total mortality and cardiovascular disease in men in the usa. Journal of Epidemiology & Community Health **50**(3), 245–251 (1996)
19. Lamb, A., et al.: Separating fact from fear: Tracking flu infections on twitter. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 789–795 (2013)
20. Mensah, G.A., Brown, D.W.: An overview of cardiovascular disease burden in the united states. Health affairs **26**(1), 38–48 (2007)
21. Naslund, J.A., et al.: How people with serious mental illness use smartphones, mobile apps, and social media. Psychiatric rehabilitation journal **39**(4), 364 (2016)
22. Paul, M., Dredze, M.: You are what you tweet: Analyzing twitter for public health. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 5, pp. 265–272 (2011)
23. Paul, M.J., Dredze, M.: A model for mining public health topics from twitter. Health **11**(16-16), 1 (2012)
24. Quincey, E.D., Kostkova, P.: Early warning and outbreak detection using social networking websites: The potential of twitter. In: International conference on electronic healthcare. pp. 21–24. Springer (2009)
25. Sinnenberg, L., et al.: Twitter as a potential data source for cardiovascular disease research. JAMA cardiology **1**(9), 1032–1036 (2016)
26. Skaik, R., Inkpen, D.: Using twitter social media for depression detection in the canadian population. In: 2020 3rd Artificial Intelligence and Cloud Computing Conference. pp. 109–114 (2020)
27. Su, K., et al.: Forecasting influenza activity using self-adaptive ai model and multi-source data in chongqing, china. EBioMedicine **47**, 284–292 (2019)