Issues in
Informing Science +
Information Technology

An Official Publication
of the Informing Science Institute
InformingScience.org

IISIT.org

Volume 18, 2021

Machine Learning-based Flu Forecasting Study Using the Official Data from the Centers for Disease Control and Prevention and Twitter Data

Ali Wahid — Colorado Technical University, Colorado Springs, CO, USA — Ali.Wahid@alumni.ctuonline.edu

Steven H. Munkeby — Colorado Technical University, Colorado Springs, CO, USA — Smunkeby@coloradotech.edu

Samuel Sambasivam* — Woodbury University, Burbank, CA, USA — Samuel.Sambasivam@Woodbury.edu

* Corresponding author

Abstract

Aim/Purpose — In the United States, the Centers for Disease Control and Prevention (CDC) tracks the disease activity using data collected from medical practices on a weekly basis. Collection of data by CDC from medical practices on a weekly basis leads to a lag time of approximately 2 weeks before any viable action can be planned. The 2-week delay problem was addressed in the study by creating machine learning models to predict flu outbreak.

Background — The 2-week delay problem was addressed in the study by correlation of the flu trends identified from Twitter data and official flu data from the Centers for Disease Control and Prevention (CDC) in combination with creating a machine learning model using both data sources to predict flu outbreak.

Methodology — A quantitative correlational study was performed using a quasi-experimental design. Flu trends from the CDC portal and tweets with mention of flu and influenza from the state of Georgia were used over a period of 22 weeks from December 29, 2019 to May 30, 2020 for this study.

Contribution — This research contributed to the body of knowledge by using a simple bag-of-word method for sentiment analysis followed by the combination of CDC and Twitter data to generate a flu prediction model with higher accuracy than using CDC data only.

Accepting Editor Eli Cohen │ Received: December 23, 2020 │ Revised: April 28, 2021 │
Accepted: June 1, 2021.
Cite as: Wahid, A., Munkeby, S. H., & Sambasivam, S. (2021). Machine learning-based flu forecasting study using the official data from the Centers for Disease Control and Prevention and Twitter data. *Issues in Informing Science and Information Technology, 18,* 63-81. https://doi.org/10.28945/4796

Findings

The study found that (a) there is no correlation between official flu data from CDC and tweets with mention of flu and (b) there is an improvement in the performance of a flu forecasting model based on a machine learning algorithm using both official flu data from CDC and tweets with mention of flu.

Recommendations for Practitioners

In this study, it was found that there was no correlation between the official flu data from the CDC and the count of tweets with mention of flu, which is why tweets alone should be used with caution to predict a flu outbreak. Based on the findings of this study, social media data can be used as an additional variable to improve the accuracy of flu prediction models. It is also found that fourth order polynomial and support vector regression models offered the best accuracy of flu prediction models.

Recommendations for Researchers

Open-source data, such as Twitter feed, can be mined for useful intelligence benefiting society. Machine learning-based prediction models can be improved by adding open-source data to the primary data set.

Impact on Society

Key implication of this study for practitioners in the field were to use social media postings to identify neighborhoods and geographic locations affected by seasonal outbreak, such as influenza, which would help reduce the spread of the disease and ultimately lead to containment. Based on the findings of this study, social media data will help health authorities in detecting seasonal outbreaks earlier than just using official CDC channels of disease and illness reporting from physicians and labs thus, empowering health officials to plan their responses swiftly and allocate their resources optimally for the most affected areas.

Future Research

A future researcher could use more complex deep learning algorithms, such as Artificial Neural Networks and Recurrent Neural Networks, to evaluate the accuracy of flu outbreak prediction models as compared to the regression models used in this study. A future researcher could apply other sentiment analysis techniques, such as natural language processing and deep learning techniques, to identify context-sensitive emotion, concept extraction, and sarcasm detection for the identification of self-reporting flu tweets. A future researcher could expand the scope by continuously collecting tweets on a public cloud and applying big data applications, such as Hadoop and MapReduce, to perform predictions using several months of historical data or even years for a larger geographical area.

Keywords

CDC, correlation, flu trends, machine learning, regression analysis

INTRODUCTION

In recent years, the use of social media platforms, such as Facebook and Twitter, has increased significantly to report news, events, sentiments, and other types of information (Comito et al., 2017). Real-time events and breaking news are shared across the world using social media. Events can be correlated to geographic location for further understanding and analysis (Comito et al., 2017). At the same time, there is a growing problem of fake information for malicious purposes, sometimes by humans after hiding their identities, and others by bots or computers using machine learning methods (Walt & Eloff, 2018). Due to the prevalence of fake information over the Internet, it is therefore necessary to validate the information for accuracy (Buntain & Golbeck, 2017).

By using big data analytical techniques on social media data, several useful trends and insights can be generated for commercial and public benefits (Byrd et al., 2016). One potential trend is to mine the tweets for insights into health trends, specifically seasonal outbreaks, such as influenza, in the given neighborhood. Such insights can lead to more focused and timely public health initiatives, which will help reduce the spread of the disease and ultimately containment (Byrd et al., 2016). Lee at al. (2017) presented the comparison of prediction accuracy between influenza triggers from Twitter and CDC versus Google Flu trends for the current week with 95% accuracy and forecasted for the following week with 89% accuracy. Traditionally, the Centers for Disease Control and Prevention (CDC) relies on lab results and doctor visits, which is why their method has an intrinsic delay of 2 weeks. In order to reduce delay in getting visibility of seasonal outbreaks, it was therefore valuable to conduct research comparing the findings of mined social media data (tweets) from a given geographic area, such as the state of Georgia, for seasonal outbreaks, such as flu versus trends from the standard body (i.e., CDC for validation and accuracy calculation). Furthermore, the data from social media was added as an independent variable along with CDC's historical data to improve the accuracy of flu prediction using machine learning algorithms. Several machine learning-based prediction models of flu were developed using regression algorithms namely, linear regression, polynomial regression, locally weighted smoothing (LOESS) regression, support vector regression, and time series, while the performance of the model was evaluated using statistical metrics, such as mean absolute error (MAE), root mean squared error (RMSE), R-Squared, and adjusted R-Squared.

PROBLEM STATEMENT

In the United States, the Centers for Disease Control and Prevention (CDC) tracks the disease activity using data collected from medical practices on a weekly basis. Collection of data by CDC from medical practices on a weekly basis leads to a lag time of approximately 2 weeks before any viable action can be planned. If insights from social media are significant and valid enough to trigger a public campaign, more focused and timely public health initiatives can be taken for a given neighborhood. This will result in a valuable social benefit by reducing CDC's 2 weeks of lag time. The 2-weeks delay problem was addressed in the study by performing the correlation of the flu trends identified from Twitter data and official flu data from the Centers for Disease Control and Prevention (CDC) in combination with creating a machine learning model using both data sources to predict flu outbreaks which, according to existing literature, has not been identified (CDC, 2020; Twitter, n.d.). The seasonal influenza outbreaks cause about 250,000 to 500,000 deaths worldwide (Lee et al., 2017). Early detection of flu outbreaks can result in a quick response, such as flu-shot campaigns, in the given geographic location, which will help in reducing the spread of disease and, in some cases, save lives (Lee et al., 2017). Such insights can lead to more focused and timely public health initiatives, which will help in reducing the spread and ultimately, containment (Byrd et al., 2016). Machine learning algorithms have been applied in several other industries; for instance, Wei et al. (2015) performed the prediction of stock prices using the stock indicator and public sentiments analysis from blogs. A study by Byrd et al. (2016) identified the influenza outbreak in Canadian cities by processing tweets. Frameworks developed by Lee et al. (2017) used CDC and tweets to predict influenza outbreaks by applying machine learning techniques.

This study contributed the body of knowledge by correlating the flu trends identified from Twitter data and official flu data from Centers for Disease Control and Prevention (CDC) and created a regression algorithm-based machine learning model using both data sources to predict the flu outbreak for the state of Georgia. This study used tweets gathered over a 22-week period from the state of Georgia only. Due to wide variability in regional flu level data per CDC (2020), this was more effective to trigger localized, more focused and timely public health initiatives, by predicting outbreaks using machine learning-based regression algorithms.

PURPOSE

The purpose of this quantitative correlational study employing quasi-experimental design was the correlation of the flu trends identified from Twitter data and official flu data from the Centers for

Disease Control and Prevention (CDC) in combination with creating a machine learning model using both data sources to predict flu outbreak. Statistical correlation of two data sources helped in determining whether insights from social media alone were significant and valid enough to trigger more focused and timely public health initiatives for a given geographic area. The research hypothesis was validated against actual CDC (2020) data for accuracy using statistical measures.

SIGNIFICANCE OF THE STUDY

Based on an extensive literature review of sentiment analysis in a general context and sentiment analysis, specifically in healthcare about the use of machine learning techniques, the subject highlighted the opportunity to conduct further research. For this research, the Twitter feed was gathered from Georgia, and mining was performed on tweets with the mention of flu and influenza to determine the likelihood of an outbreak by creating weekly trends. Processed data were then compared against the flu activity and weekly surveillance report generated by the CDC for statistical correlation and performed machine learning to predict flu outbreaks while computing statistical metrics, such as mean absolute error (MAE), root mean squared error (RMSE), R-Squared, and adjusted R-Squared. Furthermore, social media data were added as an independent variable along with CDC's historical data to improve the flu predictions in Georgia using machine learning algorithms. If social media data correlated well with CDC data, then social media data alone could be used to trigger marketing and advertising campaigns for flu shots as well as other public health initiatives.

LITERATURE REVIEW

MACHINE LEARNING

Machine learning has been in existence for several decades. Alan Turing pioneered and invented the intelligent machine in 1939 while supporting the British military in World War II (Evans & Yang, 2009; Turing, 1939). Computers have evolved a long way since then, and machines can now process the data without human intervention. Machine learning is a process of identifying patterns, performing classifications, making associations, and recognizing characters (Choi et al., 2019). The objective of machine learning is to improve the prediction of the system by using training data or old historical data (Das et al., 2017).

Machine learning algorithms are generally grouped under supervised learning, unsupervised learning, deep learning, and artificial neural networks. Supervised learning is a machine learning technique where the algorithm requires training data before predicting the test data (Zvarevashe & Olugbara, 2018). Unsupervised learning is a machine learning technique where algorithms do not require training data and learn by observation. Machine learning algorithms with multiple layers of learning from data sets and used to generate results are referred to as deep learning algorithms (Alshammari & Al-Mansour, 2019). An artificial neural network is modeled as a human brain with a network of neurons (Bayrak et al., 2019). These neurons perform certain mathematical functions on input, while outputs may go to another neuron until the final output is generated.

DATA MINING

Data mining is defined as the extraction of useful information from a large data set, often referred to as a knowledge discovery database (Santhana & Geetha, 2019). Some of the main types of data mining are classification, clustering, regression, and association rule. Clustering is a machine learning technique of grouping similar characteristics data into groups (Choi et al., 2019). All groups are represented by data points of similar characteristics. Within unsupervised learning algorithms, clustering is one of the techniques to group similar objects together without any prior information. On the other hand, classification is a supervised learning algorithm used to classify objects based on known examples referred to as training data set groups (Choi et al., 2019). Regression is a statistical technique to determine the relationship between dependent and independent variables. The dependent variable can be single or multiple, thus called univariate or multivariate regression, respectively (Kavitha et al., 2016). Association mining is a technique used to detect the rules explaining the occurrence of items within the occurrence of another set of items (Duarte & Julia, 2016). These are if-then-else statements used to identify the relationship between unrelated data variables in a dataset (Thilina et al., 2016).

Machine learning-based clustering algorithms are broken into five types: partitioning, hierarchical, density-based, grid-based, and model-based clustering algorithms (Rehioui & Idrissi, 2019). After pre-processing and feature extraction, the data are split into 90%-10%; where the 90% part is used to train the machine learning model, and thus referred to as the training phase (Ranjit et al., 2018). After the training phase, the remaining 10% of the data are used to test the machine learning model, and thus referred to as the testing phase (Ranjit et al., 2018).

Das et al. (2017) defined feature selection and reduction as an effort to pick the important features and remove the unnecessary ones from the dataset for machine learning. For instance, variance threshold is a feature selection technique, which excludes the features with the same values across all samples from the dataset (Lavanya & Mallappa, 2017). Laplacian score is another feature selection technique, which picks the features with most relevance using Eigen maps and locality preserving maps (Lavanya & Mallappa, 2017). For data extraction using text characteristics, the traditional Term Frequency-Inverse Document Frequency (TF-IDF) algorithm works by sorting the text-characteristic words per weights in the descending order (Yang, 2017).

Key Performance Metrics

For regression machine learning models, performance is measured using statistical metrics, such as mean absolute error (MAE), root mean squared error (RMSE), R-Squared, and adjusted R-Squared (Kassambara, 2018). Mean absolute error is the mean of all absolute errors for all predicted values (Kavitha et al., 2016). Root mean squared error is the square root of all mean squared errors. R-Squared is the proportion of variation in the outcome that is explained by the predictor variable. Adjusted R-Squared adjusts the R-Squared for having too many variables in the mode (Kassambara, 2018).

Sentiment Analysis

Sentiment analysis or opinion mining is a technique under natural language processing (NLP) used to understand customer and business opinions, reviews, and trends (Ikoro et al., 2018; Tanuja et al., 2019). Opinion mining or sentiment analysis are defined as the identification of people's opinions and sentiments on a given topic (Kisan et al., 2016). Sentiment analysis on tweets is difficult because of the 140-character limit as well as the usage of emoticons and symbols to share one's emotions, thus requiring significant pre-processing (Kisan et al., 2016). Walha et al. (2016) defined opinion mining as the process of identifying the opinion of people using machine learning techniques. The process to classify the text as positive, negative, or neutral is referred to as sentiment analysis (Walha et al., 2016). Similarly, Kuhamanee et al. (2017) established the definition of text mining as a process of discovering information as forms, patterns, or trends, which are hidden in original text based on statistics and mathematics.

Sentiment Analysis in Health Care

The analysis of the sentiments of social media users concerning hospitals and other medical environments is on the rise. Chaudhary and Naaz (2017), for example, developed a system to determine the reviews of hospitals by collecting tweets and analyzing them using Hadoop and R. The intent was to identify certain aspects of a given hospital based on patient comments and reviews by performing social media analysis (Chaudhary & Naaz, 2017).

A system proposed by Lee et al. (2017) used the Centers for Disease Control and Prevention (CDC) and Twitter's data feeds to predict the influenza outbreaks for the current and next week with 95% and 89% accuracy, respectively, which was comparable or better than the Google flu trend (GFT). Traditional methods used by CDC require the gathering of lab results and doctor's office visits to trend influenza, which causes a delay of 2 weeks. Google flu trend (GFT) made use of search queries to identify possible flu outbreaks in the user's location (Lee et al., 2017). Their proposed model made use of CDC and Twitter data to perform machine learning using multi-layer perceptron with back propagation algorithms to predict current and future outbreaks with high accuracy, 2 to 3 weeks faster than the traditional flu surveillance system (Lee et al., 2017).

A complete framework was presented by Byrd et al. (2016) to identify the spread of influenza in the given geographic locations, which were Ottawa, Toronto, Syracuse, and Montreal, by processing the Twitter feed using natural language processing techniques and finally, by visually showing the locations with the most tweets of flu signifying the spread of an outbreak in the given neighborhood. Their proposed method could be useful in detecting the spread of seasonal outbreaks using general consumer data instead of using lab results to launch a potential flu-shot campaign by the government (Byrd et al., 2016). Using Google mapping, Byrd et al. limited the collection of the Twitter feed to the above-mentioned cities. Keywords, such as sick, flu, or influenza, were used to filter the dataset further. Cleansing was performed by removing the stop words, hyperlinks, non-letters, and symbols. The Stanford CoreNLP algorithm, which is the best natural language processing algorithm with high accuracy, was used to tag each tweet with positive, negative, or neutral polarity. Results were displayed on the map by randomizing the location of tweets within the boundaries of four selected cities (Byrd et al., 2016).

Hidalgo (2018) conceptualized a disease prediction classification model using big data and machine learning algorithms over medical dental imaging data. Similarly, Nieto-Chaupis (2019) performed the machine learning prediction using the Weiner series to forecast the spread of flu in a city. Another study by Jung and Tonguz (2017) was conducted to analyze the sentiments of users reflected from tweets to highlight their health situation, especially weight loss. Lavanya and Mallappa (2017) collected tweets that mentioned certain types of cancer and identified the tweets that were most frequently mentioned using the k-means algorithm (which is a simple centroid-based algorithm).

Chinese social media analysis was performed by Yang et al. (2014) to identify trends of the flu disease using user postings. The proposed model was able to predict flu outbreaks 5 days ahead of the Chinese Influenza Center using social media microblogging sites with location and time information (Yang et al., 2014). Another framework was developed by Yeruva et al. (2017) to perform sentiment analysis on tweets and classify food sentiments and food types. A map was generated using the sentiments from food likings and food types, such as healthy and unhealthy, on top of CDC's obesity prevalence map for correlation. Liking of unhealthy food correlated with high obesity states and vice versa (Yeruva et al., 2017).

In summary, researchers have successfully used sentiment analysis of social media postings to identify disease outbreaks, side effects and effectiveness of prescription drugs, hospital ratings, and so on. Various machine learning algorithms, such as neural networks, classification algorithms, and series analysis were used to predict disease outbreaks for a given geographic location.

RESEARCH METHOD AND DESIGN

Qualitative research is usually employed to study human behavior and habits through exploratory, case study, and phenomenology designs (Shuttleworth, 2008). A quantitative research method is used in studies that are related to numerical data, statistical analysis, and to explain a phenomenon (Babbie, 2010). Data sets involved in quantitative studies are numerical, which are used to test the hypothesis. Per Creswell (2014), quantitative research is performed to test the hypothesis or to test the theory, establish an estimation of prevalence, or indicate the incidence of the phenomenon. This study involved the correlation and prediction of numerical data from two sources using statistical measures and machine learning algorithms, respectively. In this study, the research method or strategy of inquiry for the research was the quantitative correlational approach using a quasi-experimental design where social media data (i.e., Twitter) were collected, analyzed, and a research hypothesis was validated against official flu data from the CDC for correlational analysis using statistical measures.

RESEARCH QUESTIONS AND HYPOTHESES

The research questions for this research were as follows:

RQ1. Is there a correlation between flu outbreaks identified using Twitter data and official data from the CDC?

RQ2. What is the improvement in the accuracy of the CDC's prediction of influenza outbreaks by adding social media data as an independent variable to the machine learning algorithm?

RQ3. What type of machine learning algorithm offers the highest accuracy to predict influenza outbreaks?

The null and alternative hypotheses for this research were as follows:

Null hypothesis ($H1_0$). There is no correlation between flu data from CDC and Twitter data.

Alternative hypothesis ($H1_A$). There is a correlation between flu data from CDC and Twitter data.

Null hypothesis ($H2_0$). There is no improvement in the accuracy of the CDC's prediction of influenza outbreak by adding social media data as an independent variable to the machine learning algorithm.

Alternative hypothesis ($H2_A$). There is an improvement in the accuracy of the CDC's prediction of influenza outbreak by adding social media data as an independent variable to the machine learning algorithm.

Null hypothesis ($H3_0$). There is no machine learning algorithm that offers high accuracy to predict influenza outbreaks.

Alternative hypothesis ($H3_A$). There is a machine learning algorithm that offers high accuracy to predict influenza outbreaks.

DATA COLLECTION

Two data sets were used in this study; one of them was the official influenza repository of the CDC, while the other was publicly available tweets using Twitter's API. Weekly reported flu counts of data of 22 weeks from December 29, 2019 through May 30, 2020 by medical providers were downloaded from the CDC portal for the state of Georgia (CDC, 2020). Data from CDC was in a ready-to-use format with flu counts in one column and week information in another.

For the second data source, free Twitter login credentials were used to setup a free Twitter developer account. The Twitter developer account helped in generating the secret keys and access tokens, which were used to authenticate Twitter access and retrieve tweets using their API. Tweepy is an open-source package used to retrieve tweets using Twitter's API via secret keys and access tokens generated from the Twitter developer account. Python code was then used to retrieve tweets with keywords of flu or influenza using Georgia coordinates and a radial boundary, while filtering out retweets to reduce the duplication for the period of 22 weeks from December 29, 2019 through May 30, 2020 (Twitter, n.d.). Tweets collected date, time, location, and tweet message data.

Using the location information, only tweets with mention of Georgia were used while the remaining tweets were discarded. The total counts of tweets with mention of flu or influenza from the state of

Georgia were 37,225. Tweets were then processed using the Microsoft Excel based Visual Basic macro to convert tweets into lower case and alphanumeric characters only, which filtered out emoticons, symbols, and characters. R packages, namely plyr and stringr were used to manipulate strings by storing tweets in a vector for sentiment analysis. Preprocessing of tweets was performed in R, which removed hyperlinks, numbers, and blank spaces at the beginning and end of tweets. Tweets were then split into words using R. Using a bag-of-words approach, which was based on a list of positive and negative words from Hu and Liu (2004), each tweet was scored as a positive, neutral, or negative perspective. Using the list of personal pronouns from Alkouz et al. (2019), each tweet was then flagged as either a self-reporting or a non-self-reporting tweet. Tweets that were negatively scored and flagged as self-reported depicted the personal pain, agony, and sufferings from flu, thus only counts of those tweets were used for further correlation with CDC's official flu data. Counts of negatively scored and self-reported tweets were then broken into weekly counts and aligned with weekly dates of CDC flu counts.

INSTRUMENTATION

IBM's SPSS was used to perform correlation testing using CDC's flu data and weekly counts of tweets. Two-sample independent t-tests were used to test the significance between the weekly number of flu cases from the CDC data and the weekly number of negatively scored tweets. Pearson's correlation was also computed to determine the association between flu cases from the CDC data and the Twitter data. Finally, machine learning-based prediction models using regression algorithms were built using R to forecast the flu outbreak. The performance of the models was measured using statistical metrics, including RMSE, MAE, MSE, MAPE, R-Squared, and adjusted R-Squared.

DATA ANALYSIS

The following data analysis steps were used for this study:

STEP 1. Apply a two-sample independent t-test when the comparison between the two independent groups is desired (Boslaugh & Watters, 2008). In this study, a two-sample independent t-test was used to test the significance between the number of flu cases recorded from the official CDC data and the number of tweets. Besides the group statistics that showed the number of samples, mean, standard deviation, and standard error mean, an independent t-test checked whether the equal variance assumption was met or not using Levene's test for significance. If the significance was greater than 0.05, this would mean there was no statistically significant difference between the number of flu cases recorded at the CDC and the counts of tweets with mention of flu or influenza.

STEP 2. Pearson's correlation was used to determine the change in one variable that results from a change in another variable for a parametric data set (Boslaugh & Watters, 2008). In this study, Pearson's correlation was computed to determine the association between the number of flu cases from the CDC data and the Twitter data set. Pearson's correlation and a 2-tailed significance test helped in testing the first null hypothesis ($H1_0$), which verified the correlation between the number of flu cases recorded at the CDC data and the counts of tweets with mention of flu or influenza. Per Cohen's standard, if Pearson's correlation was greater than 0, but less than 0.1, then there was no correlation; thus, the first null hypothesis ($H1_0$) would fail to reject. If Pearson's correlation was greater than 0.1, then there was some correlation; thus, the first null hypothesis ($H1_0$) would be rejected (Boslaugh & Watters, 2008).

STEP 3. The weekly CDC data were broken into training (80%) and testing (20%) for machine learning (Dangeti, 2017).

STEP 4. A prediction of flu was performed using regression algorithms including linear regression, polynomial regression, locally weighted smoothing (LOESS) regression, support vector regression, and time series.

Step 5. Twitter data were then added as an independent variable to evaluate the improvement in prediction using regression algorithms listed in Step 4, if any. If prediction accuracy improved, then the second null hypothesis ($H2_0$) would be rejected. The machine learning algorithm with the highest accuracy to predict influenza outbreak would reject the third null hypothesis ($H3_0$).

The performance of the machine learning algorithms in Steps 4 and 5 was measured using statistical metrics, including mean absolute error (MAE), root mean squared error (RMSE), R-Squared, and adjusted R-Squared (Kassambara, 2018). Mean absolute error is the mean of all absolute errors for all predicted values (Kavitha et al., 2016). Root mean squared error is the square root of all mean squared errors. R-Squared is the proportion of variation in the outcome that is explained by the predictor variable. Adjusted R-Squared adjusts the R-Squared for having too many variables in the mode (Kassambara, 2018).

Findings

Two-sample independent t-tests between the number of flu cases from official CDC data and the number of self-reported flu tweets showed the value of F as 23.809, which is the test statistics of Levene's test and significance with a p-value of 0.000016. This led to the conclusion that the variance in counts of flu cases from official CDC data were significantly different than that of counts of self-reported flu cases on Twitter. With significantly different variances between the two data sources, the t-test for equality of means showed the value of t as 4.9777, which is the computed test statistic with a value of df as 24.976, which is the degrees of freedom with the corresponding significance p-value of 0.000040. Mean difference between the two samples was 3042.909, while standard error difference of the test statistic was 611.352. The 95% confidence interval of the difference was at the lower bound of 1783.745 and the upper bound of 4302.073. Based on the results of the independent sample t-test, it was concluded that the mean of flu cases from the CDC and the self-reported tweets was significantly different. Findings of two-sample independent t-test are tabulated in Table 1.

Table 1: Independent samples t-test of weekly flu counts from CDC and counts of self-reported flu tweets for the state of Georgia

INDEPENDENT SAMPLES TEST				
			Cases	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		23.81	
	Sig.		0.000016	
t-test for Equality of Means	t		4.98	4.98
	df		42	24.98
	Sig. (2-tailed)		0.000011	0.000040
	Mean Difference		3042.9	3042.9
	Std. Error Difference		611.4	611.4
	95% Confidence Interval of the Difference	Lower	1809.2	1783.7
		Upper	4276.7	4302.1

Pearson's correlation of the CDC flu counts with the self-reported flu counts from Twitter for the state of Georgia during the period of 22 weeks was 0.024 with a significance p-value of 0.914 for a two-tailed test. Similarly, Pearson's correlation of self-reported flu counts from Twitter with the CDC flu counts was 0.024 with a significance p-value of 0.914 for a two-tailed test based on 22 weekly observations. Based on the results of Pearson's correlation, the CDC flu counts and the self-reported flu cases from Twitter for the state of Georgia over the period of 22 weeks had no correlation, and thus failed to reject the first null hypothesis ($H1_0$), which stated that there was no correlation between the flu data from the CDC and the Twitter data. Findings of two-sample independent t-test are tabulated in Table 2.

Table 2: Pearson correlation of weekly flu counts from CDC and counts of self-reported flu tweets for the State of Georgia

CORRELATIONS			
		CDC Flu Counts	# Self-Reported Flu Tweets
CDC Flu Counts	Pearson Correlation	1	0.024
	Sig. (2-tailed)		0.914
	Sum of Squares and Cross-products	157616357.5	1191572.4
	Covariance	7505540.8	56741.5
	N	22	22
# Self-Reported Flu Tweets	Pearson Correlation	0.024	1
	Sig. (2-tailed)	0.914	
	Sum of Squares and Cross-products	1191572.4	15056465.1
	Covariance	56741.5	716974.5
	N	22	22

For the second and third null hypotheses tests, the machine learning-based regression algorithms were applied on the CDC flu data. Later, self-reported flu counts from tweets were added as the independent variable to predict flu and gauge whether flu prediction improves with the addition of self-reported flu cases from Twitter. Flu predictions were performed using nine machine learning models, namely linear regression, second order polynomial regression, third order polynomial regression, fourth order polynomial regression, LOESS with span 0.5, LOESS with span 0.6, LOESS with span 0.8, support vector regression, and time series. Modeling was performed using weekly flu counts from the CDC only, and later by using weekly flu counts from the CDC as well as the self-reported flu cases from tweets. The performance of the flu prediction models improved with the addition of self-reported flu tweets for linear regression, polynomial regression, LOESS with span 0.6, support vector regression, and time series. The performance flu prediction models degraded with the addition of self-reported flu tweets for the LOESS with span 0.5 and LOESS with span 0.8. Findings of machine learning models are presented in Figure 1 and Table 3.

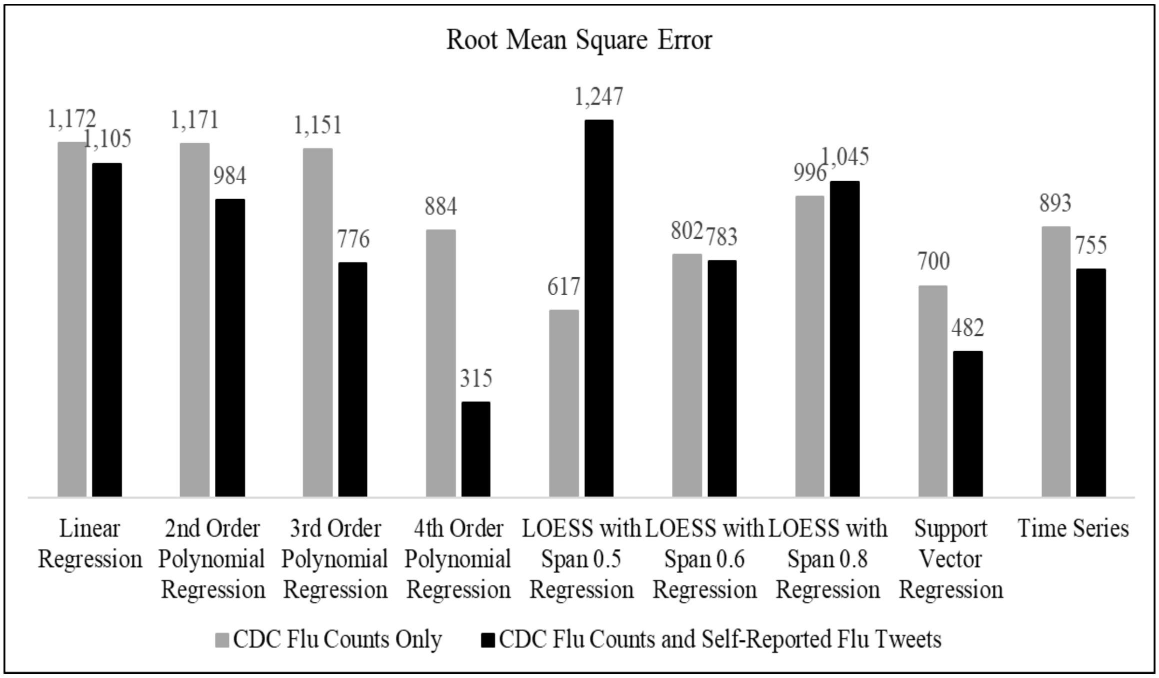

Figure 1: Comparison of root mean square error of machine learning models

Table 3: Statistical results of flu prediction using various machine learning models

Inputs	Machine Learning Algorithm	Root Mean Square Error	Mean Absolute Error	R Squared	Adjusted R Squared	Mean Squared Error	Mean Absolute Percentage Error
CDC flu counts only	Support Vector Regression	700.2	521.9	0.94	0.93	490315	15.9
	LOESS With Span 0.8 Regression	995.8	825.6	0.86	0.86	991604	26.1
	LOESS With Span 0.6 Regression	801.8	630.4	0.91	0.91	642932	16.7
	LOESS With Span 0.5 Regression	617.0	478.7	0.95	0.94	380713	12.5
	4th Order Polynomial Regression	883.6	706.9	0.89	0.89	780773	24.8
	3rd Order Polynomial Regression	1150.8	952.0	0.82	0.81	1324279	36.1
	2nd Order Polynomial Regression	1170.7	932.8	0.81	0.80	1370544	29.7

Inputs	Machine Learning Algorithm	Root Mean Square Error	Mean Absolute Error	R Squared	Adjusted R Squared	Mean Squared Error	Mean Absolute Percentage Error
	Linear Regression	1172.2	941.6	0.81	0.80	1374161	30.6
	Time Series	893.1	660.5	0.91	0.91	797570	16.4
CDC Flu Counts and Self-Reported Flu Tweets	Support Vector Regression	482.0	349.6	0.97	0.97	232310	11.3
	LOESS With Span 0.8 Regression	1045.4	729.5	0.86	0.85	1092944	16.3
	LOESS With Span 0.6 Regression	782.8	568.7	0.92	0.92	612741	13.4
	LOESS With Span 0.5 Regression	1246.6	676.3	0.81	0.79	1553960	14.2
	4th Order Polynomial Regression	315.2	229.4	0.99	0.96	99338	7.6
	3rd Order Polynomial Regression	775.7	591.4	0.92	0.85	601750	18.9
	2nd Order Polynomial Regression	983.9	797.7	0.86	0.82	968121	27.8
	Linear Regression	1104.7	939.2	0.83	0.81	1220436	32.4
	Time Series	754.9	523.9	0.97	0.97	569938	9.9

The performance of flu prediction models using statistical measures RMSE, MAE, R-Squared, adjusted R-Squared, MSE, and MAPE showed improvement for machine learning models with the addition of self-reported flu tweets, but for two models, namely LOESS with span 0.5 and LOESS with span 0.8. Based on statistical measures, it was concluded that the second null hypothesis ($H2_0$) was rejected because there was an improvement in the accuracy of flu prediction by adding social media data as an independent variable to the machine learning algorithms. The fourth order polynomial regression algorithm followed by the support vector regression models offered the best flu prediction based on statistical measures RMSE, MAE, R-Squared, adjusted R-Squared, MSE, and MAPE. Based on statistical measures, it was concluded that the third null hypothesis ($H3_0$) was rejected because there was a machine learning algorithm, which was the 4th order polynomial regression model that offered the highest accuracy to predict an influenza outbreak.

DISCUSSION

Improvement in the performance of the influenza forecast with the use of tweets in this study agreed with the framework of Lee et al. (2017), where the researchers also found that prediction of influenza outbreaks from Twitter along with the CDC data using neural networks for the entire United States had higher accuracy than the flu prediction based on Google flu trends. The influenza forecast using tweets in this study agreed with the study of Byrd et al. (2016), where the researchers identified the influenza outbreak in Canadian cities by processing tweets. Key implications of this study for practitioners in the field was to use the social media postings to identify neighborhoods and geographic locations affected by seasonal outbreaks, such as influenza. Insights from the social media data and the official CDC data could be used to predict influenza outbreaks with higher accuracy and trigger localized public health initiatives, which would help reduce the spread of the disease and ultimately lead to containment. The study highlighted the adoption of social media data as a secondary health source that could bring awareness among the public quickly regarding seasonal outbreaks; thus, resulting in an improved health posture of society. Based on the findings of this study, social media data will help health authorities in detecting seasonal outbreaks earlier than just using official CDC channels of disease and illness reporting; thus, empowering health officials to plan their responses swiftly. The study pointed to the phenomenon that prevalent usage of social media platforms can help in identifying personal and behavioral traits, which can be used for the benefit of society.

PRACTICAL IMPLICATIONS OF FINDINGS

In the United States, the Centers for Disease Control and Prevention (CDC) tracks the disease activity using data collected from medical practices on a weekly basis. Collection of data by CDC from medical practices on a weekly basis leads to a lag time of approximately 2 weeks before any viable action can be planned. If insights from social media are significant and valid enough to trigger more focused and localized public health initiatives for a given neighborhood, this will result in a valuable social benefit by reducing CDC's 2 weeks of lag time. The 2-weeks delay problem was addressed in this study by first performing the correlation of the flu trends identified from Twitter data and official flu data from the Centers for Disease Control and Prevention (CDC). Consideration was then given to the simultaneous creation of a machine learning model using both data sources to predict flu outbreaks (CDC, 2020; Twitter, n.d.).

The collection of data by the CDC from medical practices on a weekly basis leads to a lag time of approximately 2 weeks before any viable action can be planned. In this study, flu prediction models based on machine learning algorithms were built first by using the CDC data only. The performance of the models was measured using statistical metrics, including RMSE, MAE, MSE, MAPE, R-Squared, and adjusted R-Squared. Performance of the flu prediction models showed improvement across all statistical metrics with the addition of self-reported flu tweets as an independent variable to the machine learning algorithms.

Key implications of this study for practitioners in the field is to use social media postings to identify neighborhoods and geographic locations affected by seasonal outbreaks, such as influenza. Insights from the social media data and the official CDC data together could be used to predict influenza outbreaks with higher accuracy and trigger focused public health initiatives, which would help reduce the spread of the disease and ultimately lead to containment. The study highlighted that the adoption of social media data as a secondary health source could bring awareness among the public quickly regarding seasonal outbreaks; thus, resulting in an improved health posture of society. Based on the findings of this study, social media data will help health authorities in detecting seasonal outbreaks earlier than just using the official CDC channels of disease and illness reporting; thus, empowering health officials to plan their responses swiftly. The study pointed to the phenomenon that prevalent usage of social media platforms can help in identifying personal and behavioral traits, which can be used for the benefit of society.

RECOMMENDATIONS FOR PRACTITIONERS

Based on the findings of the (a) absence of correlation between official flu data from CDC and tweets with mention of flu and (b) improvement in the performance of flu forecasting models based on a machine learning algorithm using both official flu data from CDC and tweets with mention of flu, several recommendations are made for practical application.

Recommendation 1

In this study, it was found that there was no correlation between the official flu data from the CDC and the count of tweets with mention of flu, which is why tweets alone should be used with caution to predict a flu outbreak. However, social media can be used to quickly identify neighborhoods or geographic locations affected by the flu based on high social media activity using keywords of flu or influenza.

Recommendation 2

Based on the findings of this study, the forecast of flu outbreaks should be performed using both the official CDC data and the tweets with mention of flu using machine learning algorithms for higher accuracy of prediction based on lower RMSE, MSE, MAE, and MAPE versus the models using the CDC data only. It is therefore recommended that social media data be used as an additional variable to improve the accuracy of prediction models.

Recommendation 3

Based on this study, regression-based machine learning algorithms, namely fourth order polynomial models and support vector models should be used to perform the prediction of flu due to lower RMSE, MSE, MAE, and MAPE versus using other regression models. It is therefore recommended that fourth order polynomial and support vector regression models be used for improved accuracy of prediction models.

Recommendations for Further Research

Based on the findings of the (a) absence of correlation between official flu data from CDC and tweets with mention of flu and (b) improvement in the performance of flu forecasting model based on machine learning algorithm using both official flu data from the CDC and tweets with mention of flu, several recommendations are made for future research.

Future Research Recommendation 1

In this study, regression-based machine learning algorithms were used to predict flu outbreaks. A future researcher could use more complex deep learning algorithms, such as artificial neural networks and recurrent neural networks, to evaluate the accuracy of flu outbreak prediction models as compared to the regression models used in this study.

Future Research Recommendation 2

In this study, sentiment analysis of tweets was performed using the bag-of-words technique and personal pronouns to identify the self-reporting flu tweets. A future researcher could apply other sentiment analysis techniques, such as natural language processing and deep learning techniques, to identify context-sensitive emotion, concept extraction, and sarcasm detection for the identification of self-reporting flu tweets.

Future Research Recommendation 3

In this study, prediction models were built using only 22 weeks of data for the state of Georgia. A future researcher could expand the scope by continuously collecting tweets on a public cloud and applying big data applications such Hadoop and MapReduce to perform predictions using several months of historical data or even years for a larger geographical area.

Conclusion

The problem addressed in the study was the correlation of the flu trends identified from Twitter data and official flu data from the CDC in combination with creating a machine learning model using both data sources to predict flu outbreak (CDC, 2020; Twitter, n.d.). The collection of data by the CDC from medical practices on a weekly basis leads to a lag time of approximately 2 weeks before any viable action can be planned. The purpose of this quantitative correlational study employing a quasi-experimental design was to correlate the flu trends identified from Twitter data and official flu data from the CDC to create a machine learning model using both data sources to predict flu outbreaks.

The conclusions for the study based on the findings were (a) there is no correlation between official flu data from CDC and tweets with mention of flu and (b) there is an improvement in the performance of flu forecasting models based on machine learning algorithms using both official flu data from CDC and tweets with mention of flu. Improvement in the performance of influenza forecast with the use of tweets in this study agreed with the framework of Lee et al. (2017), where the researchers also found that predictions using influenza triggers from Twitter with CDC data had higher accuracy versus Google flu trends. Influenza forecasts using tweets in this study agreed with the study of Byrd et al. (2016), where the researchers identified the influenza outbreak in Canadian cities by processing tweets.

Future Research

Key implication of this study for practitioners in the field was to use social media postings to identify neighborhoods and geographic locations affected by seasonal outbreak, such as influenza. Insights from social media data and official CDC data together could be used to predict influenza outbreaks with higher accuracy and trigger focused health initiatives, which would help reduce the spread of the disease and ultimately lead to containment. Similar approaches could be used to track the prevalence of other diseases, such as heart problems and respiratory issues. The study highlighted that the adoption of social media data as a secondary health source could bring awareness among the public regarding seasonal outbreak thus, resulting in an improved health posture of society. Public awareness advisories can be issued based on health trends identified from social media platforms, which will enable the public to take precautionary measures thus, protecting them from getting sick. Based on the findings of this study, social media data will help health authorities in detecting seasonal outbreaks earlier than just using official CDC channels of disease and illness reporting from physicians and labs thus, empowering health officials to plan their responses swiftly and allocate their resources optimally for the most affected areas. The study pointed to the phenomenon that prevalent usage of social media platforms can help in identifying personal and behavioral traits, which can be used for the benefit of the entire society. Similar to the prediction of influenza outbreaks in this study, an increase in respiratory illnesses reported in social media due to pollen or bad air quality in the given location can be used to advise the public to use a face mask or take anti-allergy medicines.

References

Alkouz, B., Aghbari, Z., & Abawajy, J. (2019). Tweetfluenza: Predicting flu threads from Twitter data. *Big Data Mining and Analytics*, *2*(4), 273–287. https://doi.org/10.26599/BDMA.2019.9020012

Alshammari, N., & AlMansour, A. (2019, May). State of the art review on Twitter sentiment analysis. *Proceedings of the 2nd International Conference on Computer Applications & Information Security (ICCAIS)* (pp. 1–8). Riyadh, Saudi Arabia: IEEE. https://doi.org/10.1109/CAIS.2019.8769465

Babbie, E. R. (2010). *The practice of social research (2nd edition)*. Wadsworth Cengage.

Boslaugh, S., & Watters, P. (2008). *Statistics in a nutshell (1st edition)*. O'Reilly.

Buntain, C., & Golbeck, J. (2017, November). Automatically identifying fake news in popular Twitter threads. *Proceedings of the IEEE International Conference on Smart Cloud* (pp. 208–215). New York, NY, USA: IEEE. https://doi.org/10.1109/SmartCloud.2017.40

Bayrak, E., Kirci, P., & Ensari, T. (2019, April). Comparison of machine learning methods for breast cancer diagnosis. *Proceedings of the Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1–3). Istanbul, Turkey: IEEE. https://doi.org/10.1109/EBBT.2019.8741990

Byrd, K., Mansurov, A., & Baysal, O. (2016, May). Mining Twitter data for influenza detection and surveillance. *Proceedings of the 38th International Workshop on Software Engineering in Healthcare Systems (ICSE '16)* (pp. 43–49). Austin, TX, USA: ACM. https://doi.org/10.1145/2897683.2897693

Centers for Disease Control and Prevention (CDC) (2020, April 21). *FluSight: Flu forecasting.* https://www.cdc.gov/flu/weekly/flusight/index.html

Chaudhary, S., & Naaz, S. (2017, October). Use of big data in computational epidemiology for public health surveillance. *Proceedings of the International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)* (pp. 150–155). Gurgaon, India: IEEE. https://doi.org/10.1109/IC3TSN.2017.8284467

Choi, H. W., Qureshi, N. M. F., & Shin, D. R. (2019, February). Analysis of electricity consumption at home using K-means clustering algorithm. *Proceedings of the 21st International Conference on Advanced Communications Technology (ICACT)* (pp. 639–643). PyeongChang, South Korea: IEEE. https://doi.org/10.23919/ICACT.2019.8701981

Comito, C., Falcone, D., & Talia, D. (2017, October). A peak detection method to uncover events from social media. *Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 460–467). Tokyo, Japan: IEEE. https://doi.org/10.1109/DSAA.2017.69

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches (4th edition).* Sage Publications.

Dangeti, P. (2017). *Statistics for machine learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R (1st edition).* Packt.

Das, M. K., Padhy, B., & Mishra, B. K. (2017, January). A review opinion mining and sentiment classification. *Proceedings of the International Conference on Inventive Systems and Control (ICISC)* (pp. 1–3). Coimbatore, India: IEEE. https://doi.org/10.1109/ICISC.2017.8068637

Duarte, V. A. R., & Julia, R. M. S. (2016, December). Improving the state space representation through association rule. *Proceedings of the 15th International Conference on Machine Learning and Applications* (pp. 931–934). Anaheim, CA, USA: IEEE. https://doi.org/10.1109/ICMLA.2016.0167

Evans, J. R. G., & Yang, S. (2009). Solid freeforming and combinatorial research. *Tsinghua Science and Technology, 14*(S1), 94–99. https://doi.org/10.1016/S1007-0214(09)70074-4

Hidalgo, J. (2018, March). *Exploring the big data and machine learning framing concepts for a predictive classification model.* Doctoral Dissertation. Colorado Springs, CO, USA: Colorado Technical University.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)* (pp. 168–177). Seattle, WA, USA: ACM https://doi.org/10.1145/1014052.1014073

Ikoro, V., Sharmina, M., Malik, K., & Batista-Navarro, R. (2018, October). Analyzing sentiments expressed on Twitter by UK energy company consumers. *Proceedings of the 5th International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 95–98). Valencia, Spain: IEEE. https://doi.org/10.1109/SNAMS.2018.8554619

Jung, K. S., & Tonguz, O. K. (2017, July). Using social networks to predict changes in health. *Proceedings of the 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (pp. 12–13). Bratislava: IEEE. https://doi.org/10.1109/SMAP.2017.8022659

Kassambara, A. (2018). *Machine learning essentials: Practical guide in R (1st edition).* Sthda.

Kavitha, S., Varuna, S., & Ramya, R. (2016, November). A comparative analysis on linear regression and support vector regression. *Proceedings of the 2016 Online International Conference on Green Engineering & Technologies (IC-GET)* (pp. 1–5). Coimbatore, India: IEEE. https://doi.org/10.1109/GET.2016.7916627

Kisan, H. S., Kisan, H. A., & Suresh, A. P. (2016, December). Collective intelligence & sentimental analysis of Twitter data by using Stanford NLP libraries with Software as a Service (SaaS). *Proceedings of the International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-4). Chennai, India: IEEE. https://doi.org/10.1109/ICCIC.2016.7919697

Kuhamanee, T., Talmongkol, N., Chaisuriyakul, K., San-Um, W., Pongpisuttinun, N., & Pongyupinpanich, S. (2017, July). Sentiment analysis of foreign tourists to Bangkok using data mining through online social network. *Proceedings of the 15th International Conference on Industrial Informatics (INDIN)* (pp. 1068–1073). Emden, Germany: IEEE. https://doi.org/10.1109/INDIN.2017.8104921

Lavanya, P. G., & Mallappa, S. (2017, September). Automatic summarization and visualization of healthcare tweets. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1557–1563). Udupi, India: IEEE. https://doi.org/10.1109/ICACCI.2017.8126063

Lee, K., Agrawal, A., & Choudhary, A. (2017, August). Forecasting influenza levels using real-time social media streams. *Proceedings of the International Conference on Healthcare Informatics (ICHI)* (pp. 409–414). Park City, UT, USA: IEEE. https://doi.org/10.1109/ICHI.2017.68

Nieto-Chaupis, H. (2019, January). Face to face with next flu pandemic with a Wiener-series-based machine learning: Fast decision to tackle rapid spread. *Proceedings of the 9th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 654–658). Las Vegas, NV, USA: IEEE. https://doi.org/10.1109/CCWC.2019.8666474

Ranjit, S., Shrestha, S., Subedi, S., & Shakya, S. (2018, October). Foreign rate exchange prediction using neural network and sentiment analysis. *Proceedings of the International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 1173–1177). Greater Noida, India: IEEE. https://doi.org/10.1109/ICACCCN.2018.8748819

Rehioui, H., & Idrissi, A. (2019). New clustering algorithms for Twitter sentiment analysis. *IEEE Systems Journal*, *14*(1), 530–537. https://doi.org/10.1109/JSYST.2019.2912759

Santhana, K. J., & Geetha, S. (2019, April). Prediction of heart disease using machine learning algorithms. *Proceedings of the 1st International Conference on Innovations in Information and Communication Technology (ICIICT)* (pp. 1–5). Chennai, India: IEEE. https://doi.org/10.1109/ICIICT1.2019.8741465

Shuttleworth, M. (2008, March 07). *Quantitative research design*. Explorable.com. https://explorable.com/quantitative-research-design

Tanuja, U., Gururaj, H. L., & Janhavi, V. (2019, January). An exploratory analysis on data features and analysis techniques in social networks. *Proceedings of the 11th International Conference on Communication Systems and Networks (COMSNETS)* (pp. 535–537). Bengaluru, India: IEEE. https://doi.org/10.1109/COMSNETS.2019.8711472

Thilina, A., Attanayake, S., Samarakoon, S., Nawodya, D., Rupasinghe, L., Pathirage, N., Edirisinghe, T., & Krishnadeva, K. (2016, December). Intruder detection using deep learning and association rule mining. *Proceedings of the International Conference on Computer and Information Technology (CIT)* (pp. 615–620). Nadi, Fiji: IEEE. https://doi.org/10.1109/CIT.2016.69

Turing, A. M. (1939). Systems of logic based on ordinals. *Proceedings of London Mathematical Society*, *s2-45*(1), 161–228. https://doi.org/10.1112/plms/s2-45.1.161

Twitter (n.d.). *About public and protected tweets*. https://help.Twitter.com/en/safety-and-security/public-and-protected-tweets

Walha, A., Ghozzi, F., & Gargouri, F. (2016, November). A lexicon approach to multidimensional analysis of tweets opinion. *Proceedings of the 13th IEEE/ACS International Conference of Computer Systems and Applications (AICCSA)* (pp. 1–8). Agadir, Morocco: IEEE. https://doi.org/10.1109/AICCSA.2016.7945704

Walt, E., & Eloff, J. (2018). Using machine learning to detect fake identities: Bots vs human. *IEEE Access*, *6*, 6540–6549. https://doi.org/10.1109/ACCESS.2018.2796018

Wei, G., Zhang, W., & Zhou, L. (2015, July). Stock trends prediction combining the public opinion analysis. *Proceedings of the International Conference on Logistics, Informatics and Service Sciences (LISS)* (pp. 1–6). Barcelona, Spain: IEEE. https://doi.org/10.1109/LISS.2015.7369692

Yang, N., Cui, X., Hu, C., Zhu, W., & Yang, C. (2014, October). Chinese social media analysis for disease surveillance. *Proceedings of the International Conference on Identification, Information and Knowledge in the Internet of Things* (pp. 17–21). Beijing, China: IEEE. https://doi.org/10.1109/IIKI.2014.11

Yang, Y. (2017, October). Research and realization of internet public opinion analysis based on improved TF-IDF algorithm. *Proceedings of the 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)* (pp. 80–83). Anyang, China: IEEE. https://doi.org/10.1109/DCABES.2017.24

Yeruva, V. K., Junaid, S., & Lee, Y. (2017, November). Exploring social contextual influences on healthy eating using big data analytics. *Proceedings of the International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1507–1514). Kansas City, MO, USA: IEEE. https://doi.org/10.1109/BIBM.2017.8217885

Zvarevashe, K., & Olugbara, O. O. (2018, March). A framework for sentiment analysis with opinion mining of hotel reviews. *Proceedings of the Conference on Information Communications Technology and Society (ICTAS)* (pp. 1–4). Durban, South Africa: IEEE. https://doi.org/10.1109/ICTAS.2018.8368746

AUTHORS

Ali Wahid is associated with AT&T Mobility in the capacity of Principal Radio Access Network Engineer and as an adjunct faculty with Department of Computer Science at Grand Canyon University, USA. He has 16+ years of experience in the telecommunication industry to design and optimize cellular wireless networks while working in Pakistan, Saudi Arabia and United States with Nortel, Ericsson, and Nokia. Ali co-authored US Patent No. 2017-1063 to identify small cells using machine learning. Ali completed his Bachelors in Electronics Engineering from NED University of Engineering and Technology - Pakistan, Master's in Communications Engineering from Nanyang Technological University – Singapore and Doctorate in Computer Science from Colorado Technical University – USA.

Dr. Steve Munkeby has worked and managed high technology projects and programs for the federal government while in industry for 30 years. Expertise includes serving as an Army Infantry Officer followed by work on the Space Shuttle and payloads for the Shuttle (software developer), earth observing spacecraft (software manager), and on robotic/unmanned ground vehicles (program manager). In addition, he has concurrently served 15 years of progressive doctoral teaching roles including leading doctoral faculty instructing core/concentration courses and as dissertation director of dissertation chairs/committee members. He has a B.S. in Computer Science from the University of Montana, M.S. in Systems Management from the University of Southern California, and a Doctorate in Management (specializing in organizational leadership) from the University of Phoenix.

Dr. Samuel Sambasivam is Chair and Professor of Computer Science Data Analytics at Woodbury University, Burbank, CA. He is Chair Emeritus and Professor Emeritus of Computer Science at Azusa Pacific University. He served as a Distinguished Visiting Professor of Computer Science at the United States Air Force Academy in Colorado Springs, Colorado for two years. In addition, he has concurrently served 13 years of progressive doctoral teaching roles at Colorado Technical University (CTU) including chair of doctoral programs, lead computer science doctoral faculty instructing core/concentration computer science courses and as dissertation chair/committee member. His research interests include Cybersecurity, Big Data Analytics, Optimization Methods, Expert Systems, Client/Server Applications, Database Systems, and Genetic Algorithms. He has conducted extensive research, written for publications, and delivered presentations in Computer Science, data structures, and Mathematics. Dr. Samuel Sambasivam earned his Ph.D. in Mathematics/Computer Science from Moscow State University, a Master of Science in Computer Science with Honors from Western Michigan University, Pre-PhD in Mathematics/Computer Science from Indian Institute of Technology (IIT) Delhi, a Master of Science Education in Mathematics with Honors from Mysore University (NCERT-Delhi), and a Bachelor of Science in Mathematics/Physics/Chemistry with Honors from the University Madras (Chennai). He is a voting senior member of the ACM.