Twitter sentiment reason mining Framework to identify major problems in the USA Healthcare Industry.

E.M. Rasika Chamara Edirisinghe

A dissertation submitted in partial fulfilment of the requirement for Master of Science degree in Business Analytics

Department of Computing Informatics Institute of Technology, Sri Lanka

in collaboration with

Robert Gordon University, Scotland

2023

ROBERT GORDON UNIVERSITY ABERDEEN

SCHOOL OF COMPUTING SCIENCE AND DIGITAL MEDIA

MSc Big Data Analytics

Student Name: Rasika Chamara Edirisinghe Mudiyanselage	Matriculation Number:
Supervisor: Dinesh Asanka	Second Marker:
Project Title: Twitter sentiment reason mining Framework to identify major problems in the USA Healthcare Industry.	
	Start Date: 2022 September
	Submission Date: 2023 August

CONSENT

I agree ■
I do not agree □

That the University shall be entitled to use any results, materials or other outcomes arising from my project work for the purposes of non-commercial teaching and research, including collaboration.

DECLARATION

I confirm:

- **That the work contained in this document has been composed solely by myself and that I have not made use of any unauthorised assistance.**
- **That the work has not been accepted in any previous application for a degree.**
- **All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.**

Student Signature:	Date Signed: 08.16.2023

Abstract

This research presents a comprehensive study on USA healthcare by applying machine learning and natural language processing techniques on Twitter data. The study aims to gain valuable insights into public sentiment towards healthcare-related issues and break down the overall twitter discourse on the USA healthcare domain. The framework presented involved multiple approaches, including sentiment analysis, sentiment spike detection, clustering, keyword extraction, topic modeling, and textual association each contributing to a deeper understanding of the complexities within the healthcare domain.

First, a custom Snscrape algorithm was utilized to extract 4,925,984 tweets regarding selected keywords related to USA healthcare over a period of 24 months from January 2021 to December 2022. In the subsequent stage, a comprehensive textual preprocessing pipeline consisting of URL removal, emoji handling, non-alpha numeric character handling, case folding, language detection, language translation, Stopword removal, tokenization, and lemmatization was implemented on the collected tweets to obtain a standardized dataset.

In the sentiment analysis, in the pre-trained category, three models were utilized: AFINN, TextBlob, and VADER. The pre-trained models provided an initial understanding of the sentiment expressed in the collected tweets. In addition to the pre-trained models, this research also explores the use of various custom trained models for sentiment analysis, such as Decision Trees, Logistic Regression, Random Forest, RNN_TDFVec, CNN_TDFVec, and DNN_TDFVec. Furthermore, TextBlob, VADER and AFINN were also custom trained in this study. By employing a range of pretrained and trained models, the aim of this step was to identify the most suitable approach for sentiment analysis in the context of USA healthcare-related tweets. Both trained and pretrained models were tested with various metrics such as Accuracy, Precision, Recall, and F1 Score. Furthermore, the neural network-based models were also evaluated regarding the training and validation loss for both train and test data, and the most accurate model was selected to be incorporated into the framework.

Two sentiment spike detection approaches were employed to identify significant shifts in public sentiment. The first approach uses Python's Matplotlib library to represent sentiment values graphically and identifies spikes based on the disparity between positive and negative sentiments. However, this approach faces computational challenges with larger datasets. To address this, the second approach creates a more efficient data subset, resampling the data to daily frequency, and computing Z-scores to identify spikes based on deviations from the mean sentiment. Both approaches offer valuable insights into sudden shifts in public sentiment, allowing for further investigation into the underlying reasons for these fluctuations. The second approach was incorporated into the framework and the sentiment spikes were identified, and the data related to each spike was moved forward within the framework.

Next, several clustering models were evaluated, including K-Means, Agglomerative, Birch, and MeanShift, to determine the most suited keyword extraction algorithm to be used within the framework Additionally, the effectiveness of multiple vectorization models was tested, such as

CountVectorizer, TF-IDF, and Word2Vec, to represent textual data numerically. The effectiveness of all the clustering and vectorization model combinations were evaluated using Silhouette Score, Davies Bouldin Score, and the Calinski Harabasz Score and the most suited model was implemented within the framework. The elbow method was employed to customize the number of clusters for each detected sentiment spike. The elbow method helped identify the optimal number of clusters by assessing the variance explained by each clustering mode. The top keywords were extracted per each sentiment spike and brought forward within the framework.

In the next stage, topic modeling was applied to the spike data to extract key topics for each sentiment spike. Three popular topic modeling models are evaluated: Latent Dirichlet Allocation, Latent Semantic Analysis, and Non-Negative Matrix Factorization (NMF) regarding metrics such as Perplexity score, Coherence CV, Coherence U Mass, Coherence C UCI, Topic Diversity, and Topic Dominance Score. The top model was selected and was used to extract topics related to each sentiment spike, which was also brought forward within the framework.

Next, TF-IDF scores were calculated per each topic-keyword combination for each spike dataset to determine the optimum terms related to each sentiment spike.

In the subsequent framework stage, each spike dataset was analyzed to extract the top hashtags associated with the relevant tweets. In the domain of sentiment mining on Twitter, hashtags are pivotal in revealing the underlying subjects and themes of tweets.

In the final stage of the search term compilation, both the top TF-IDF topic-keyword combinations and the top hashtags were utilized to compile an optimum set of terms which captures the essence of each sentiment spike, and these terms were used to mine news articles related to each spike.

This was done using a custom algorithm utilizing the googlenews and beautifulsoup python libraries. Furthermore, the date of the sentiment spike and the geolocation was also utilized to target the article mining to the relevant time frame and the relevant geolocation of interest, being the United States. The search terms were used in an iterative sliding window mechanism to extract all top news articles related to the sentiment spike. Per each mined news article, the title, link, publication date, source, description, author, and region were extracted and saved for the final analysis of the framework.

The extracted news article titles and descriptions were compiled into a single corpus per each article and were textually preprocessed to remove any noise or bias within the data. Next three textual association scores, Cosine Similarity, Sequence Similarity, and Levenshtein Distance were used to calculate an overall association score between each news article and the search terms and the top 10 news articles per each spike was presented for manual inspection as the outcome of the sentiment mining framework.

By integrating these multiple approaches, our research contributes to a comprehensive and data-driven understanding of public sentiment towards the USA healthcare system. The findings have significant implications for stakeholders, including policymakers, healthcare organizations, and marketers. Data-driven insights can support evidence-based decision-making and foster positive