IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Research Article)

Sentiment analysis using healthcare data

Ibidapo Ayobami Adeyiola [1,*], Taiwo Michael Ayeni [2] and Elijah Ayooluwa Odukoya [3]

[1] *Department of Statistics, Faculty of Science and Technology, Federal Polytechnic Ugep, Cross River, Nigeria.*
[2] *College of Professional Studies, Analytics, Northeastern University Toronto, Canada.*
[3] *Department of Statistics, Faculty of Science, Ekiti State University, Ado-Ekiti, Ekiti State, Nigeria.*

International Journal of Science and Research Archive, 2024, 13(02), 3143-3148

Publication history: Received on 08 November 2024; revised on 17 December 2024; accepted on 19 December 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2476

Abstract

Healthcare sentiment analysis focuses on diagnosing healthcare-related issues that people have discovered. To create rules and changes that could directly address patients' issues by considering their input. Machine learning techniques analyze millions of review documents and conclude them towards an efficient and accurate decision. This study trained the health data using different machine-learning algorithms: Multinomial Naïve Bayes, Random Forest, Decision Trees, K-nearest neighbor, and Support Vector Machine and compared the accuracy using different evaluation metrics. The study used a drug review dataset from the UCI machine-learning repository. It was separated so that 70% of the data was used as a training dataset for the ML models. The remaining 30% of the data forms the test dataset used to evaluate the trained ML models.

Based on the evaluation metrics, the random forest has the highest accuracy (89.4%) and R squared (0.501), and the lowest MSE (10.5) and RSME (0.324). The study concluded that the random forest classifier is the optimal model for predicting healthcare data while KNN has the lowest accuracy. It is recommended government health ministry and healthcare facilities use online health data to create policies that will directly address these public health issues, allowing patients to directly address their concerns to higher authorities without having to go through arduous procedures

Keywords: Sentiment analysis; Machine Learning; Online Health data; Social Media

1. Introduction

Technology is moving at a faster and more innovative pace these days and because of this advancement and increase in the availability of internet tools and a growth in social media platforms for personal blogs, opinion sites, and websites that evaluate products online. People now use several platforms to convey their thoughts, feelings, and ideas and take advice on them. Which have drawn interest from interested parties like clients, businesses, and governments to examine and examine these viewpoints.

Individuals use blogs and forums to talk about their ailments, symptoms, drugs, and other health-related subjects. There is also discussion on the accessibility, service, atmosphere, comfort, contentment, and other characteristics of the local healthcare facilities visited. Hearing about other people's experiences choosing a healthcare institution, taking medications, or making decisions regarding their health is often beneficial for new patients.

Medical facilities require this information to identify and treat their patients' problems. This information can help hospitals and healthcare providers better understand and handle their patients' interests and concerns. The study's strongest point is the patients' sharing of experiences covered by their sentiment analysis and passions.

[*] Corresponding author: Ibidapo Ayobami Adeyiola

Sentiment analysis studies people's perceptions of a subject and its attributes. Medical content is not only generally accessible but also freely available online, so manually evaluating this enormous amount of data is less effective.[1]

Sentiment analysis is a method used to determine whether the information gathered is neutral, negative, or positive. It consists of analyzing the feelings associated with a written work on any subject. Sentiment analysis examines a variety of viewpoints, such as those on politics, celebrities, cuisine, locations, or other subjects, to assess people's beliefs, preferences, opinions, and interests [2]. Since there is a vast number of free health-related content available online, it is impractical to manually evaluate all of this data and draw conclusions to make a quick and effective choice. With little to no assistance from users, sentiment analysis systems carry out this duty through automated procedures. Patients' opinions expressed in millions of papers across various platforms are taken into consideration using sentiment analysis. The classification of health decisions into two classes—recommended and non-recommended. [3] stated that the following are the various levels of sentiment analysis that can be used: the first level identifies each line's neutral, negative, and positive sentiment. The second level acknowledges the complete sentiment record as a single unit, whether it be positive, negative, or neutral. The next level is used when attributes are present in a post, input text, or entity and the four-level handles the social connections between various clients by utilizing graph theory

This article aims to train the health data using different machine learning techniques and compare the accuracy of the technique's prediction. Using this data, healthcare facilities and the government health ministry can formulate policies that will directly address these issues affecting the general public, enabling patients to voice their concerns to higher authorities directly without undergoing grueling procedures.

The rest of the paper is organized as follows: section (2) illustrates the related work relevant to sentiment analysis using health datasets and machine learning algorithms. Section (3) describes the different machine learning algorithms and the evaluation metrics used to evaluate the performance of the selected classifiers; section (4) presents the data analysis and the achieved results. Finally, Section 5 concludes the paper.

2. Related Work

Some empirical literature review has been done that covers various sentiment analysis, emotion recognition, handwritten/typed text analysis, and Natural language processing, Specifically,

Saranya [4] in the paper Sentiment Analysis of Healthcare Tweets using the SVM Classifier, proposed a framework for the investigation of Twitter facts by first removing the tweet using Twitter API. The preprocessed tweets were characterized and predicted using SVM grouping. Khan & Khalid [5] explored sentiment analysis for healthcare. The study highlighted the usefulness of sentiment analysis in health care. The study found that Supervised methods such as Naïve Bayes, KNearestNeigbhour, Logistic Regression, and unsupervised classification can improve sentiment health care. It concluded that supervised techniques have high accuracy but are not extendable to unknown domains while unsupervised techniques have low accuracy.

Altawaier & Tiun [6] investigated the performance of three machine learning (ML) techniques including Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) when used on Arabic sentiment analysis based on a simple extracted feature. The dataset was Arabic tweets collected from the UCI repository. The Results showed that the Decision Tree classifier outperformed the other ML by obtaining 78% of the f-measure.

Jadhav [7] examined the current and future landscape of Natural Language Processing (NLP) in healthcare, focusing on the integration and potential of generative AI and GPT-3 technologies. The study highlighted how these advanced technologies can streamline healthcare data management, enhance patient engagement, and facilitate innovative research methodologies. The paper also critically examined the challenges and limitations inherent in the application of NLP within healthcare. The research underscores the need for rigorous standards and ethical considerations in the development and implementation of NLP tools in healthcare.

Similarly, Cavalcanti & Prudêncio [8] proposed a new unsupervised and knowledge-based method for extracting aspects in drug reviews. The proposed solution is based on linguistic features, more specifically dependency paths in the syntactic tree of a review. The quality of the dependency path rules was investigated in several experiments in review corpora associated with three different diseases. Promising results were achieved compared to previous work.

3. Methods

In this section, we have presented the dataset description, the different machine-learning techniques, and the evaluation metrics. In this paper, we used a drug review dataset obtained from the UCI machine learning repository that has six parameters namely drug-id, name, condition, review, rating, and some taken diseases (Heart Attack, Depression, Diabetes, Cancer, Tuberculosis, etc). The five different machine learning techniques adopted are random forest, K-Nearest Neighbors, Decision trees, support vector machine, and multinomial navies bayes.

3.1. Machine Learning Algorithms

In machine learning, the Random Forest classifier/algorithm is a potent tree-learning method. In the training stage, it generates several Decision Trees. It uses a variety of decision trees on different dataset subsets and averages them to increase the dataset's predicted accuracy. The random forest forecasts the final output by taking the predictions from each decision tree and using the majority votes of predictions, rather than depending on only one decision tree.

The K-Nearest Neighbors (KNN) algorithm is a supervised machine-learning technique for resolving regression and classification issues. In machine learning, KNN is one of the most fundamental yet crucial categorization algorithms. It is heavily used in pattern recognition, data mining, and intrusion detection and falls within the supervised learning category.

Decision Trees (DTs), which can be thought of as a piecewise constant approximation, are a non-parametric supervised learning technique used for regression and classification. The objective of DTs is to learn basic decision rules inferred from the data features to develop a model that predicts the value of a target variable.

Based on the Bayes theorem, Multinomial Naive Bayes (MNB) is a well-liked and effective machine learning technique. It is frequently employed in text categorization jobs that require handling discrete data, such as document word counts. We will talk about and put MNB into practice in this essay.

Over time, a support vector machine was built and effectively applied to regression, classification, and outlier detection problems with nonlinear systems. However, it was initially employed to distinguish the two classes. The statistical learning theory is the foundation for this supervised parametric machine learning method. The SVM method draws a parallel line or hyperplane between the data that comprise the classes to distinguish the two classes.

3.2. Evaluation metrics

To evaluate the accuracy of the metrics, some metrics such as MSE (Mean square error), RMSE (Relative root mean squared error), R-Square, and R2 (Coefficient of determination) were used to compare the performance success of the prediction models used in the present paper.

4. Results and discussion

The results of the analysis are presented and discussed in this section. The descriptive data, including a chart and summary statistics, are displayed in Table 1.

Table 1 Summary Statistics

Statistics	**drug name**	**condition**	**Rating**
Count	215063	215063	215063
Mode	Levonorgestrel	Birth Control	Good
Lowest	Limbrel 500	Tympanostomy	Bad

Table 1 shows that Levonorgestrel is the most frequent drug name among the 215,063 unique data, while Limbrel 500 has the lowest. Tympanostromy has the lowest score for the condition, whereas birth control has the greatest score. Good is at the top, and bad is at the bottom. Additionally, Figure 1 lists the top ten medical conditions. The most common conditions among the respondents are birth control, depression, pain, anxiety, and acne, respectively.

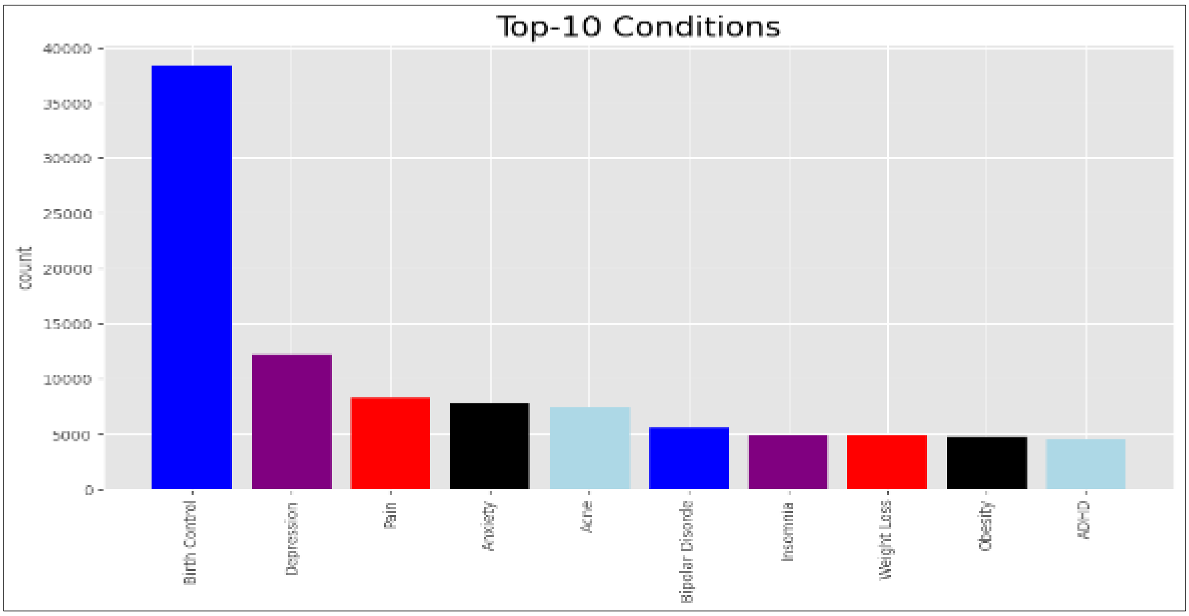

Figure 1 Top 10 Condition

Moreover, figure 2 included the top ten drug names used by the respondents. The top five drug names are, in order, Levonorgestrel, Etongestrel, Norethindrone, Nexplanon, and Norgestimate,

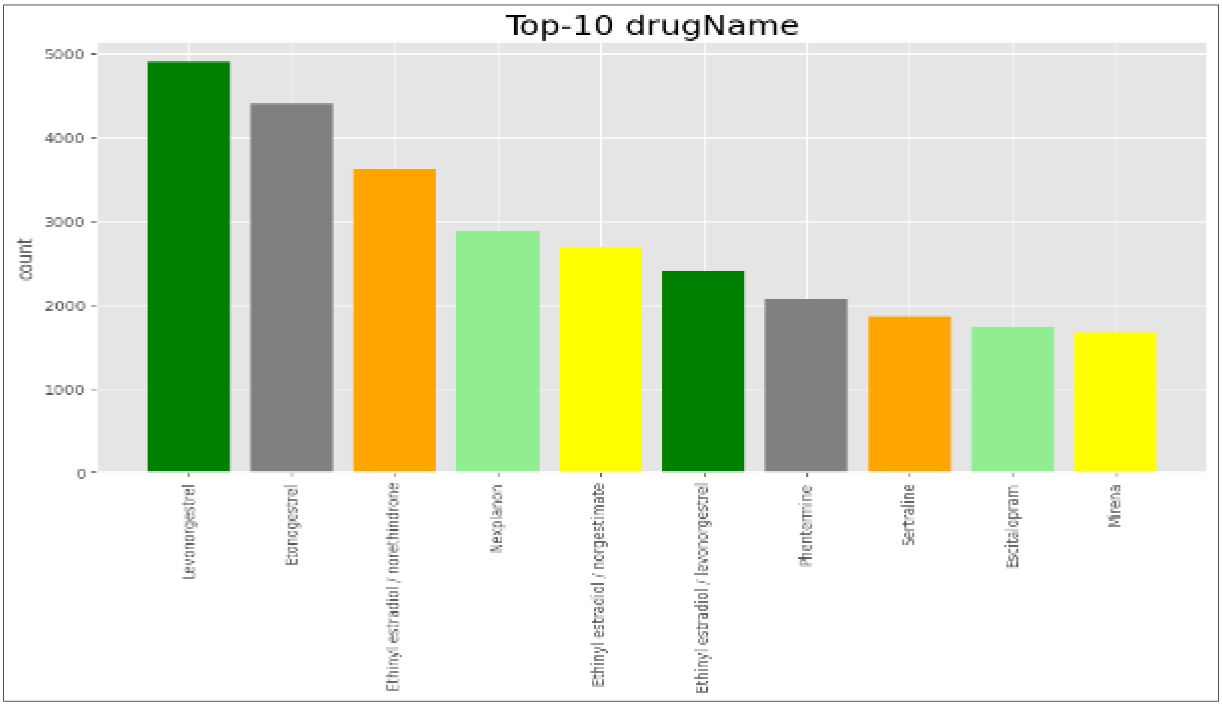

Figure 2 Top 10 drug names

Table 2 Evaluation of different Machine Learning Algorithms

Model	Accuracy	MSE	RSQUARED (R2)	RMSE
Multinomial Naive Bayes	75.5%	0.245	-0.166	0.495
K Nearest Neighbour (KNN)	71.8%	0.293	-0.396	0.541
Random Forest	89.4%	0.105	0.501	0.324
Decision Tree	71.2%	0.290	-0.382	0.539
Support Vector Machine (SVM)	83.5%	0.168	0.215	0.410

To evaluate the success of these algorithms, five different machine learning, which are frequently used in the literature, were used to train and predict the dataset. The datasets were separated so that 70% of the data was used as a training dataset, used to train the ML models. The remaining 30% of the data forms the test dataset, which is used to evaluate the trained ML models

The result in Table 2 shows the accuracy of each machine learning algorithm, the mean square error (MSE), the R-squared(R2), and the root mean square error (RMSE). Based on the classification accuracy, the Random Forest classifier is the best machine learning with an accuracy of 89.4% closely followed by the support vector machine with an accuracy of 83.5% while the decision tree has the least accuracy of 71.2%. Based on the mean square error, Random Forest with an MSE score of 0.105 has the lowest MSE score, the lowest among the techniques, while KNN has a score of 0.293. This implies that Random Forest is the optimal technique based on Mean Square Error.

Multinomial Naives, KNN, Random Forest, and Decision Trees have scores of -0.167, -0.396, 0.501, and 0.215, respectively, according to RSquared. based on this outcome. There is no doubt that Random Forest is the greatest technique because it has the highest score, 0.501. Lastly on the result of RMSE, Multinomial Naive Bayes have an RMSE of 0.495, KNN with a score of 0.541, Random Forest with a score of 0.324, Decision tree has a score of 0.539 and SVM have a score of 0.410. Based on the result, the Random Forest has the lowest RMSE.

5. Conclusion

This paper primarily uses various machine-learning algorithms to train and evaluate health data and then compares the outcomes using various evaluation measures. Based on the result of the different evaluation metrics. Random forest classifier outperforms other techniques for the considered data set. The study concludes that Random Forest is the optimal machine-learning technique for health data. The study advised the government health ministry and healthcare facilities to use online health data to create policies that will directly address these public health issues, allowing patients to directly address their concerns to higher authorities without having to go through arduous procedures.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of sentiment analysis for capturing patient experience from free-text comments posted online. Journal of medical Internet research, 15(11), e2721.

[2] Akshi Kumar and Teeja Mary Sebastian, “Sentiment Analysis on Twitter”, International Journal of Computer Science Issues, Volume 9, Issue 4, No. 3, July 2012.

[3] Kiruthika M., Sanjana Woonna, Priyanka Giri (2016). Sentiment Analysis of Twitter Data International Journal of Innovations in Engineering and Technology, Volume 6, Issue 4, April 2016.

[4] Saranya, G., Geetha, G., Meenakshi, K., & Karpagaselvi, S. (2020, December). Sentiment analysis of healthcare Tweets using SVM Classifier. In 2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS) (pp. 1-3). IEEE.

[5] Khan, M. T., & Khalid, S. (2016). Sentiment analysis for health care. In Big data: concepts, methodologies, tools, and applications (pp. 676-689). IGI Global.

[6] Altawaier, M. M., & Tiun, S. (2016). Comparison of machine learning approaches on Arabic twitter sentiment analysis. International Journal on Advanced Science, Engineering and Information Technology, 6(6), 1067-1073.

[7] Jadhav, S., Shiragapur, B., Purohit, R., & Jha, N. (2024). A Systematic Review on the Role of Sentiment Analysis in Healthcare.

[8] Cavalcanti, D. C., & Prudêncio, R. B. (2017, May). Unsupervised aspect term extraction in online drugs reviews. In The Thirtieth International Flairs Conference.