Development of a Medical Condition Prediction Model Using Natural Language Processing with K-Nearest Neighbour

[1]Bolaji A. Omodunbi, [*2]Afeez A. Soladoye, [3]Nnamdi S. Okomba, [4]Charity S. Odeyemi and [5]Mutiu O. Ayinla
[1,2, 3]Department of Computer [1]Engineering, Federal University Oye-Ekiti, Nigeria
[3]Department of Computer Engineering, Federal University Akure, Nigeria
[4]Department of Computer Science, Kwara state College of Education, Ilorin, Nigeria
{bolaji.omodunbi|Afeez.soladoye|Nnamdi.okomba}@fuoye.edu.ng|csodeyemi@futa.edu.ng |mo.ayinla@kwcoeilorin.edu.ng

Received: 16-FEB-2024; Reviewed: 12-MARCH-2024; Accepted: 18-MARCH-2024
https://dx.doi.org/10.4314/fuoyejet.v9i1.4

ORIGINAL RESEARCH

Abstract— Capturing the effect of drugs being used by patients and using this review to predict the medical ailment they are facing is a good approach to easily predict medical conditions. A lot of researchers use clinical and demographic data (risk factors) to predict diseases, the limitation of this approach is that not all the instances would have the right clinical results and there is usually missing values, low prediction accuracy, inadequately pre-processed dataset, failure to consider feature selection and un-experimentation of alternative values of K when using K-nearest neighbour. Using drug review would go a long way as their effect and symptoms as reported by the user through their review would capture relevant information needed. This study employed an open access drug review dataset to predict the medical condition, this dataset consist of training and testing split which was integrated and later split using 80-20 splitting with stratification. The dataset went through some natural language processing techniques such as lemmatization, stemming, removal of stop words, tokenization, and vectorization among others. Forward –backward feature selection technique was employed with the comments having significant effect to the prediction of the condition. K-nearest neighbour was then employed to predict the medical condition using the drug review as the feature with the condition as the target variable. Different values of nearest neighbours were used to train the model with k=1 given the best predictive average accuracy of 89% with weighted average precision of 90%. The model gave the same average accuracy of 84% when k was initialised to 3, 4, 5 and 6 respectively. Moreover, the model obtained a better result when compared with exciting systems. Therefore, with the use of artificial intelligence, medical doctors and patients can easily use drug review to predict certain medical condition using clinical predictive decision support system.

Keywords— drug reviews, natural language processing, KNN, medical condition.

1 INTRODUCTION

Technological advancement most especially the wide spread of AI application most especially in the health sector has made it possible to employ it solving many problems. Recently, the application of natural language processing known as sentiment analysis has been employed to extract meaningful information from customer review in e-commerce. This can be used for variation of purposes like recommendation of product to customer, improvement on customer's service and detection of goods quality (Anil *et al., 2018*). In health sector, customer are also leaving review for the drugs they have used for a specific medical condition, this can be explored through the incorporation and hybridization of natural language processing and machine learning techniques to make variety of prediction namely: prediction of customer's condition based on the drug review stated (Das, Badhon and Jalal, 2022).

Sentimental analysis is a field of natural language processing which give room for analysis of people's opinion to have an insight about their view or emotion about a specific matter.

*Corresponding Author: afeez.soladoye@fuoye.edu.ng

Section B- ELECTRICAL/COMPUTER ENGINEERING & RELATED SCIENCES
Can be cited as:
Omodunbi B.A, [*2]Soladoye A.A. , [3]Okomba N.S., [4]Odeyemi C. S. and [5]Ayinla M.O. (2024). Development of a Medical Condition Prediction Model Using Natural Language Processing with K-Nearest Neighbour, FUOYE Journal of Engineering and Technology (FUOYE**JET**), **9(1), 25--32.** https://dx.doi.org/10.4314/fuoyejet.v9i1.4

Natural language processing has also found its application in domains like news classification (Sharama *et al.*, 2021; Khanam *et al.*, 2021; Padalko *et al.*, 2023; Alarfaj and Khan, 2023; Mehta *et al.*, 2024), cyberbullying detection (Ali and Sayeed, 2020; Chahat *et al.*, 2021) Customer's opinion on the effect of the drug taken and how they felt or effect on them can be used to understand the nature of their disease during medical diagnosis. This approach involves the analysis of keyword presents in their review matched to a specific medical condition for easier prediction (Uddin *et al.*, 2022). Conventional medical approach involves traditional diagnosis through physical question of patient and their response (review) is used by the Medical Doctor to understand their condition. Apparently, this is quite time consuming, stressful and costly as consultation fees sometimes might be much and their might not be enough medical personnel to attend to all patients available.

As a result of this, the advancement of machine learning techniques which provide accurate, faster and easier prediction would proffer solution to the aforementioned drawbacks of the conventional medical approach. Many studies have employed several traditional machine learning and deep learning algorithm to prediction of patient's medical condition using their reviews from the drugs used, however, the major drawback of some of these studies are low prediction accuracy, inadequately pre-processed dataset, failure to consider feature selection and un-experimentation of alternative values of

K when using K-nearest neighbor. This study aim to develop a prediction system that uses the customer's drug review to predict the disease condition of patient using natural language processing and K-nearest neighbor, with consideration of major NLP pre-processing techniques so as to enhance the study's predictive performance. Moreover, studies conducted for the classification of patients' disease condition employed clinical dataset which might not be able to capture al the needed risk factors, in contrast this study employed patients' review as they would give detail description of the effect of the drug on them which would simplify the prediction of their medical condition (Gräßer, *et al.*, 2018). Many studies have been conducted on natural language processing and sentiment analysis. In recent time, some researchers have also conducted researches on the prediction of patient's medical condition using their reviews of certain drugs used for treatment. Some of these literatures and similar ones are reviewed in this section and compared to understand the state-of-art in drug reviews for prediction.

Uddin *et al.* (2022) conducted a study on the sentiment analysis of drug reviews to categorize drugs into different classes based on their effectiveness. This study employed drug review dataset obtained from UCI machine learning data repository, of which various natural language processing data pre-processing techniques were applied on the dataset such as lemmatization and tokenization, chi-squared was used for features selection. These pre-processed data was used for prediction with four machine learning algorithms namely: RF, SVM, NB and MLP as the classifiers. Random forest gave the best performance accuracy among all the classifiers employed. Similarly, Vijayaraghavan and Bas (2020) conducted a research on the sentiment analysis to detect the user's sentiment from the drug review to verify the rating given to such drug. The study also employed the same dataset as employed in their previous study. The study only considered only three condition (depression, birth control and pain) compared to the earlier study that considered, count vectorizer was used for the vectorization of the reviews. The study also used word-to-vector and different machine learning algorithms were employed for the detection namely: ANN, LSTM, GRU, SVM, RF and LR. These algorithms gave a good detection accuracies. However, this study does not consider majority of NLP techniques to improve the performance of the model.

Dinh, Chakraborty and McGaugh (2020) employed the principle of natural language processing to classify the drug reviews based on the effectiveness and side effect rating of the drug. The study used dataset obtained from Drulib.com and drugs.com, both being the widely used pharmaceutical information resource repository. The dataset is downloadable form the popular UCI machine learning data repository. This dataset is better than the earlier dataset used in the previous studies as the reviews were categorised into side effect reviews, comment reviews and benefit reviews of which in the other dataset all the reviews were not categorized.

The study was implemented in SAS® Enterprise Miner™ and data partition node was used to split the dataset into training (70%) and validation (30%). Side effect level and effectiveness of the drug were mined using clustering and text rule builder gave the best performance in predicting the target variable. Chauhan *et al.* (2021) conducted similar study to analysis the sentiment in drug review employing the same platform- SAS® Enterprise Miner™- used by the earlier study. The study employed some data preprocessing on the drug review dataset obtained from the UCI ML repository database such as conversion to lowercase, removal of punctuations, stopwords, null and numerical values, and stemming. Sentiment analysis was performed using WIT.AI- a chatboot that is capable of parsing textual or vocal messages into structured data. The study developed an interface where medicine that were rated good for a specific illness stated by the user are displayed and other diseases that could be cured using same medicine. However, the technicality behind the developed platform and implementation of the system were not well established and reported.

Das *et al.* (2021) also conducted a similar study on disease prediction from drug information provided in the users' drug reviews. The UCI ML repository dataset was also employed with removal of special character, punctuations, URLs and mentions. Label encoder was used to encode the labels and TF-IDF vectorizer was used for the vectorization. Afterwards, random forest was used for the classification and a good classification performance was obtained, Naïve Bayes and LSTM with embedding were also employed for the classification but they were all outperformed by random forest. Major pre-processing like lemmatization, stemming, removal of stop words and punctuation marks were not considered in this study, which might have caused the low performance of the models. A study conducted by Aditya and Rawat (2019) similarly employed the UCI ML repository drug review dataset to predict the customer's sentiment using the reviews on drugs provided by them. The study removed instance with missing values in the condition column and text summarization was performed by on the cleaned data which might cause removal of important keywords and instances and in-turn reduce the predictive performance of the model resulting from un-generalized instances and attributes, similarly, TF-IDF vectorizer was used for the vectorization, after which Neural network, random forest, naïve Bayes and XGBOOST were used for the sentiment prediction. Neural network gave the best performance on both rating and sentiment prediction.

All the reviewed literatures employed different conventional machine learning and deep learning algorithms for sentiment analysis of drug reviews and prediction of medical conditions. From all the aforementioned literatures none of the studies employed k-nearest neighbor as the condition prediction algorithms

https://dx.doi.org/10.4314/fuoyejet.v9i1.4 engineering.fuoye.edu.ng/journal

and most of the studies employed the UCI-ML repository drug review dataset being the only data on such subject matter. This study would also employ the dataset but with K-nearest neighbor to fill the gap of its usage.

The stages involved in this methodology includes; data acquisition, data preprocessing, prediction and performance evaluation. This is represented in block diagram as shown in Figure 1 for easier clarification and representation.

2 METHODOLOGY

`This study employed a common methodology widely used by most natural language processing problems that involves machine learning for prediction or classification.
.

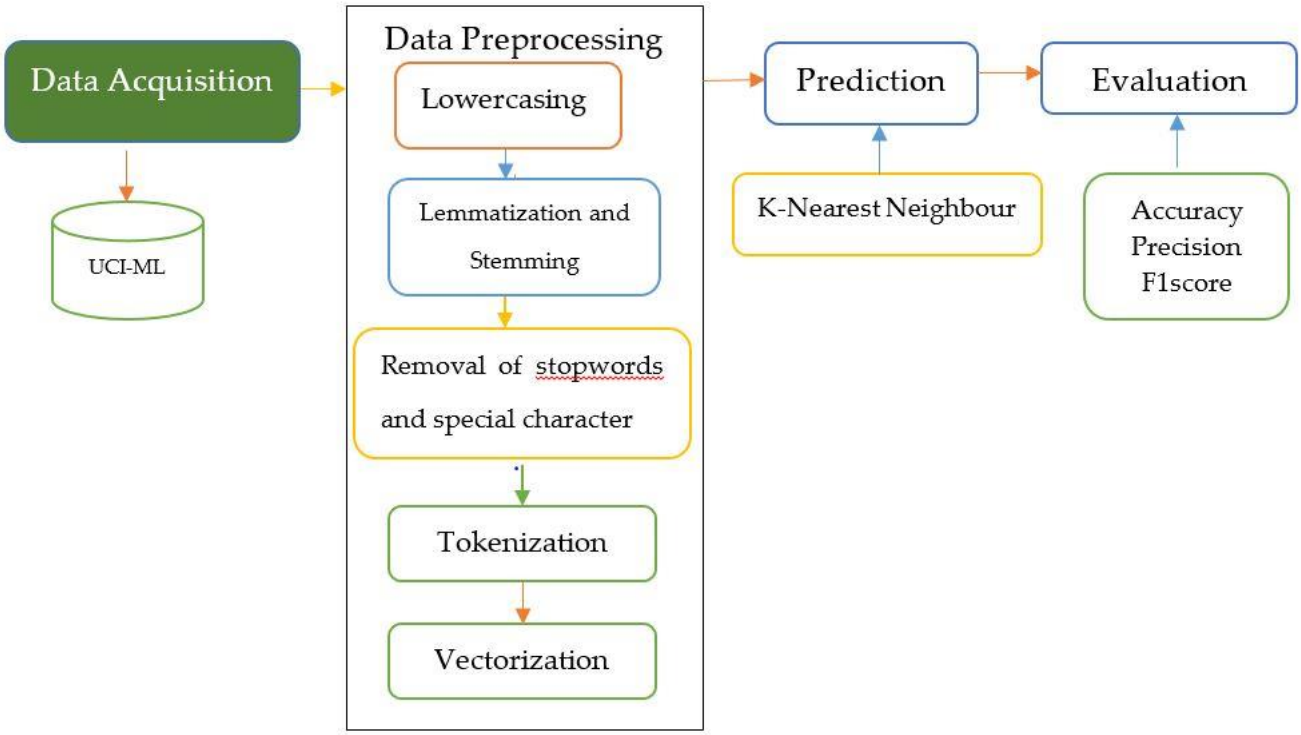

Figure 1: Overview of the research methodology

2.1 DATA ACQUISITION

The dataset used in this study was the common patients' drugs review dataset acquired from the UCI machine learning repository. This dataset consist of six (6) attributes namely: drug name, condition, review, rating, date and useful count, with a total of 215,063 instances. This dataset is the most popularly used drug review data for sentiment analysis and disease condition prediction using drug reviews. The dataset comprises of two different dataset namely "drugsComTest" for the test split which is 30% of the whole dataset and "drugsComTrain" which is the remaining 70%. The whole dataset was acquired which was later split using hold out evaluation method of 80-20. The dataset consist of common medical conditions like Birth control, Depression, Pain, Anxiety, Acne, Bipolar disorder, Obesity, Diabetes type 2, high blood pressure among other. Birth control have the highest number of instance with 38436 records, followed by depression with 12164 instances.

2.2 DATA PREPROCESSING

Natural language involves the use of some preprocessing techniques to ensure it suitability for machine learning processing. This prepeorcessing techniques helps in formatting and putting the textual data in a right format that would be implementable on machine learning model and give a good computational performance. As stated in Figure 1, some of these preprocessing techniques employed are: Vectorization, tokenization, lemmatization, stemming, and removal of punctuation, special characters and stopwords. All this would enhance the performance of the model on the data and give a good predictive performance.

https://dx.doi.org/10.4314/fuoyejet.v9i1.4 engineering.fuoye.edu.ng/journal

2.2.1 CONVERSION TO LOWERCASE

The review comprises of both upper and lower cases, which might cause inconsistency in the text. The reviews were firstly converted to lowercases such that all the letters on the review are all in lower case for uniformity. This would enable the model to easily recognize the alphabets during vectorization and they would not be considered as different character represented with different vector. For instance, the sentence I used to take…. Would change to i used to take. And other sentences like that.

2.2.2 LEMMATIZATION

Lemmatization is an important preprocessing techniques employed in natural language processing. This is the process of grouping inflected forms of a word together so they can be analyzed as a single item, identified by the word's lemma or dictionary form (Uddin *et al.,* 2022). Lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. This is a text pre-processing technique used in natural language processing (NLP) models to break a word down to its root meaning to identify similarities. To actualize this wordnet Lemmatizer was used being a very popular and actively used algorithm for word lemmatization

2.2.3 STEMMING

This is the process of removing suffixes or affixes that are added to a word. It is the reduction of a word back to its root or stem form after the inflection have been removed. With the reduction of word to their stem it gives room for the model to focus on the main word for classification and helps in accurate classification. Porter Stemmer was used for stemming in this research work. This is different from lemmatization because suffix and prefix are not removed but they are reduced to dictionary form for easier understanding and identification. Porter stemmer was also used in this regard.

2.2.4 REMOVAL OF STOPWORDS, PUNCTUATIONS AND SPECIAL CHARACTERS

Stopwords are categories of words that does not add any meaning to the body of the sentence same thing applicable to punctuations and special characters. Though in speaking stopwords are important as they add understanding to conversation and punctuation also mean a lot in speaking. However, we can do without them in language computation as they are always removed in natural language processing. Some of the commonly removed English stopwords from corpus are pronouns, auxiliary verbs, and preposition among others. The common stopwords removed from the reviews are: i, me, am, it, my, among others.

2.2.5 TOKENIZATION

This is the process of breaking textual dataset into smaller pieces like words, sentences, terms and any other syllabic elements, these smaller pieces are known as Tokens. This is sometimes the first stage in natural language processing techniques. Tokenizer breaks stream of unstructured textual data into discretized elements. Tokenizer was imported differently from the text preprocessing library.

2.2.6 VECTORIZATION

This techniques employed bag of word model to convert the textual data into their corresponding numerical format using count_vectorizer that creates Document–Term Matrix by considering the frequency of each word in the dataset and represent them as a vector. This is paramount in natural language processing as only a numerical data format can be fed to the machine learning model for manipulation.

3.2.7 K-Nearest Neighbor (K-NN)

K-NN is also a supervised learning algorithm, lazy learner though, as it is widely called as no model is learned and non-parametric algorithm, like DT it is also used for solving classification and regression problems. This algorithm makes its own prediction based on the proximity of the class label (Soladoye, 2023), as it assumes that similar and related things must exist in close proximity to each other. Equation 1 represents the Euclidean distance which is usually used to calculate the distance of these neighbors to each other. When a new instance is to be predicted, the closest to it among all the already classified instances is calculated using Euclidean distance or Manhattan distance, the class of the instance that is closest to it is the solution to the new instance we're trying to predict, as shown in Figure 2. The *k* in the K-NN implies the number of the nearest neighbor to put into consideration in the voting process (Kotsiantis *et al.*, 2006).

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{k}(xi - yi)^2} \quad (1)$$

K= Number of neighbours

x_i = training data and y_i = test data

https://dx.doi.org/10.4314/fuoyejet.v9i1.4 engineering.fuoye.edu.ng/journal

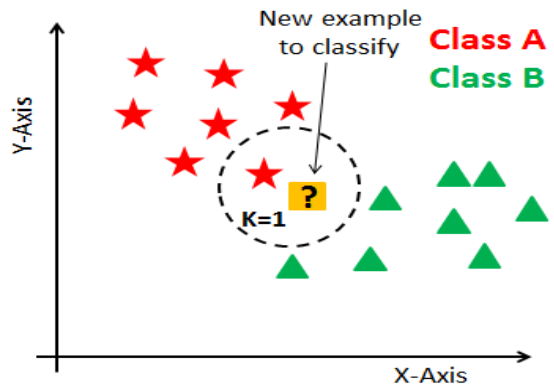

Figure 2: K- Nearest Neigbour (Ak, 2020)

This study employed different values of k ranging from 1-6, this is done to assess and compare the performance of the model with different values of nearest neighbour in order to know the number of neighbours with the best predictive performance.

2.3 Experimental Implementation

The experimentation was conducted on Google Colab, a cloud based jupyter notebook for implementing machine learning projects. This platform is quite impressive as some of machine learning libraries were already preinstalled, and possess a fast computation time compared to running the notebook on local machine. This study employed the aforementioned dataset but only four common medical conditions were considered namely: Birth control, Depression, Diabetes, Type 2 and High blood pressure. This conditions amount to a total of 57066 instances. Only drug review was the attributes used for prediction of the medical condition, because the drug name that would have been used have multiple drugs for a single condition and many conditions have same drug. This might be difficult for the model to understand as there would be misconception and inconsistency, which is why the attribute was dropped while other attributes have no relevance to the prediction of the medical conditions as well.

3.2 Performance Evaluation

The developed system's performance was evaluated using some evaluation metrics like accuracy, Sensitivity, Precision, F1 score and computational time.

i. Accuracy: This measures the overall effectiveness of the developed system and it is measured in percentage (%), as represented in Equation 2

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (2)$$

ii. Precision: It depicts the number of truth positive (positive classes) predicted that really belong to the positive class. It is given mathematically by Equation (3.)

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

iii. F1 score: This is the harmonic mean of recall and precision. It is given mathematically by equation (4)

$$\text{F1score} = \frac{2*precision*recall}{precision+recall} \quad (4)$$

3 Results and Discussion

This section reports the results obtained from the implementation of the machine learning model for the prediction of medical condition based on the patients' reviews of the drug taken. The experimental results obtained from using different values of k would be discussed and the text data would be visualized based on the target variable using word cloud function in python

3.1Textual data visualization

Visualizing the text data in a whole dataset would possibly show the frequency of occurrence of a specific word in the dataset. This would be done based on the medical conditions of the patients with their review. Word cloud in python was used to generate this word cloud based on the condition. Figures 3-6 shows the word cloud of the four conditions considered for prediction based on the reviews obtained for the customers.

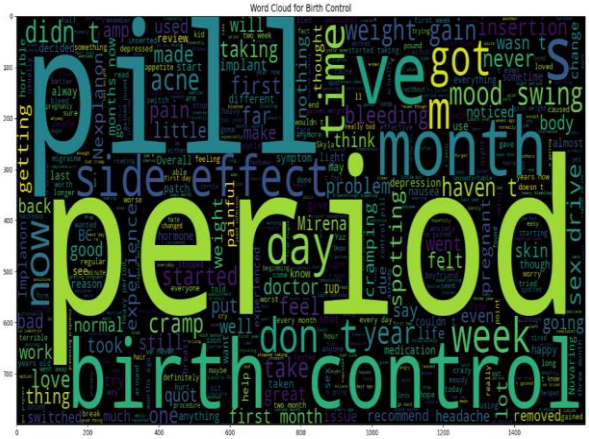

Figure 3: Word cloud for birth control condition

This shows the frequently surfaced word in the patients' reviews with respect to the birth control as the medical condition. From Figure 3, it would be observed that words like pills, period, mood swing, cramp, month and others were bold showing the level of their frequency and truly related to birth control condition. This gives an insight of the possible words available in the review even if it is not possible to explore all the reviews individually.

https://dx.doi.org/10.4314/fuoyejet.v9i1.4 engineering.fuoye.edu.ng/journal

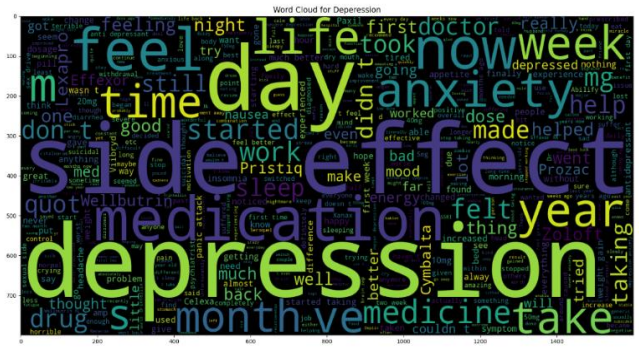

Figure 4: Word cloud for Depression condition

Similarly, the words like side effect, medication, life, night, week and other are frequent in the depression condition reviews, as these were represented in Figure 4 for clarity and easier representation. All this words are related either positively if negatively to the depression condition and express the patients' states.

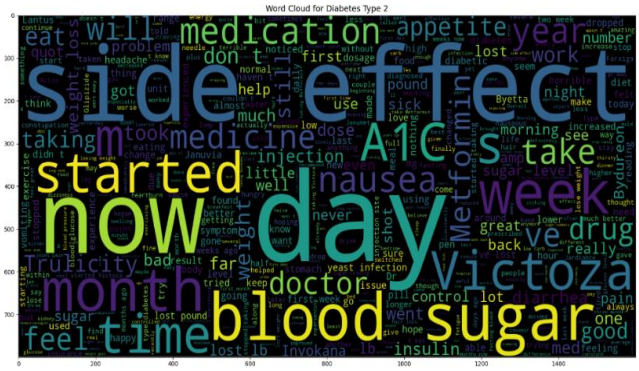

Figure 5: Word cloud for diabetes type 2 condition

Insulin is obviously captured by the word cloud likewise nausea, side effect and victoza, these are also words seen to have been present in different reviews given by different patients suffering from diabetes type 2 as shown in Figure 5.

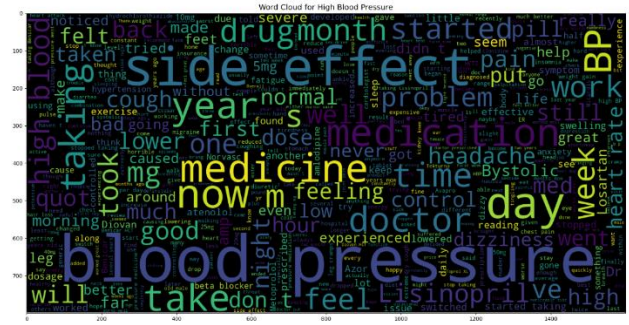

Figure 6: Word cloud for blood pressure condition

The final condition considered in the study is blood pressure with different words that surfaced in the reviews related to it captured in Figure 6. These words include BP, headache and dizziness. Word cloud is a very good text visualization tool which shows the frequency of a word.

3.3 Experimental Result

The medical condition is predicted using K-nearest Neighbor. Different values of k were used for the prediction. Table 1 shows the classification report of the model with k ranging from 1 to 6.

Table 1: Classification reports of different values of k

Classes and Metrics	Precision	Recall	F1score	Support
		K=1		
Birth control	0.97	0.93	0.95	5758
Depression	0.78	0.87	0.82	1814
Diabetes T2	0.70	0.71	0.71	511
High BP.	0.65	0.73	0.69	464
Accuracy			0.89	8547
Macro avg.	0.78	0.81	0.79	8547
Weighted avg.	0.90	0.89	0.89	8547
		K=2		
Birth control	0.94	0.95	0.94	5758
Depression	0.73	0.84	0.78	1814
Diabetes T2	0.68	0.44	0.53	511
High BP.	0.78	0.50	0.61	464
Accuracy			0.87	8547
Macro avg.	0.78	0.68	0.72	8547
Weighted avg.	0.87	0.87	0.87	8547
		K=6		
Birth control	0.95	0.90	0.92	5758
Depression	0.64	0.87	0.74	1814
Diabetes T2	0.73	0.35	0.48	511
High BP.	0.68	0.56	0.61	464
Accuracy			0.84	8547
Macro avg.	0.75	0.67	0.69	8547
Weighted avg.	0.85	0.84	0.84	8547

As shown in Table 1, three different values of k were presented namely: 1, 2 and 6. K=3, 4 and 5 were not presented because they gave the same average accuracies like k=6, however the precision weighted averages for k=6 was the best among the three that was why it's classification report was presented.

Moreover, it can be observed from the support column that the conditions does not have same number of instances which implies that the dataset was not balanced, however the imbalance nature of the dataset does not really affect the performance of the model as all the conditions were recognized and well classified. As a result of the imbalance nature of the dataset, weighted averages would be the best metrics values to consider for all the evaluation metric.

Additionally, it would be observed that k=1 gave the best predictive performance f=of 89% average with weighted average precision of 90% the highest among all values of

https://dx.doi.org/10.4314/fuoyejet.v9i1.4 engineering.fuoye.edu.ng/journal

k. The small instances of the conditions would not be said to have affected the values obtained as the classification report of k=2 the precision obtained on high blood pressure medical condition is greater than that of depression even with the fact that depression have 1814 test instances while high blood pressure has just 484 instances.

Based on these results, KNN gave the best predictive performance when k=1 and the next high performance was obtained when k=2 and after that value, the average accuracies of the model remained 84%.

3.4 Comparison with Existing System

The results obtained from this research was compared with other system that employed the same dataset to predict the medical condition or other variables in the dataset. Table 2 shows result of the existing studies and that obtained from this study.

Table 2: Comparison of the developed system with existing studies

S/N	Author	Methodology/ Algorithm	Evaluation
1	Das *et al.*, 2021	Random forest	Acc=0.83
2	Joshi and Abdelfattah (2021)	Linear SVC	Pre. =0.8832
3	Das, Badhon and Jalal (2022)	LSTM	Acc=0.81
4	This study	KNN	Acc=0.89. Pre=0.90

From the comparison result presented in Table 2, it was observed that this study outperformed all the existing studies on prediction of medical conditions using drug review with natural language processing techniques. Even the study that employed deep learning algorithm which is expected to work better on textual dataset gave a lesser average accuracy than that of the resulted obtained through this research. This shows that the study was well taken and employed all the need NLP and ML techniques that improved the performance of the predictive system.

4 Conclusion

Based on the result presented earlier in Table 1, k-NN (k=1) gave the best predictive performance on the drug review dataset acquired from UCI-ML repository to predict the medical condition the reviewer was suffering from. However, technically it might not be appropriate to opt for the solution with k=1 as it would easily predict when its nearest neighbour is just one (1) but even with 87% given with k=2 and the 84% given by k= 3, 4, 5 and 6 are good accuracies whose system can be implemented to solve real life problems and employed in real time to predict a medical condition using the drugs reviews. Moreover weighted averages of the evaluation reports were considered because the dataset used was imbalanced and also the precision was discussed being the best evaluation metrics to evaluate the performance of a machine learning model on imbalanced data and k=1 also gave the best weighted precision of 90% and 87% was obtained when k was initialized to 2. This shows that K-NN is a very good machine learning model that could be used absolutely for the prediction of medical condition using natural language processing.

The results obtained from this study was also compared with that of other studies and they were all outperformed, even that of the deep learning algorithm. Future work should explore deep learning algorithm for the prediction to compare their performance with the employed machine learning algorithms.

References

Aditya, A., and Rawat, S. (2019). Review Sentiment Analysis and Rating Prediction on Drug Review Dataset. Accessed on 12th February 2024 retrieved from https://www.researchgate.net/publication/348944498

Ak, M. F. (2020) "A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications" *Healthcare, 8(111)*

Alarja, F. K. and Khan, J. A. (2023). Deep Dive into fake news detection: Feature-Centric CLassification with Ensemble and Deep Learning Methods. *Algorithms, 16, 507*

Ali, A. and Syed, A. M. (2020). Cyberbully detection using machine learning *Pakistan Journal of Engineering and Technology (PekJET), 5(1), 45-50*

Anil, D., Vembar, A., Hiriyannaiah, S. and Srinivasa, K. G. (2018).Performance Analysis of Deep Learning Architectures for Recommendation Systems. *IEEE 25th International Conference on High Performance Computing Workshops (HiPCW), Bengaluru, India, 129-136*

Chahat, R., Ayush, A., Gnana, B., Bhuya, N. and Mukesh, P. (2021). Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques" – *10th Anniversary of Electronics: Recent Advances in Computer Science & Engineering) https://doi.org/10.3390/electronics10222810*

Chauhan, S., Bahl, V., Sengar, N. and Goel, A. (2021). Sentiment Analysis of Drug Reviews Using Wit.AI. *International Research Journal of Modernization in Engineering Technology and Science, 3(12)*

Das, S., Badhon, A. J. and Jalal, M. (2022). Predicting Effectiveness of Drug from Patient's Review. *In Proceedings of the 2nd International Conference on "Advancement in Electronics & Communication Engineering (AECE 2022) July 14-15*

Das, S., Mahata, S. K., Das, A. and Deb, K. (2021). Disease Prediction from Drug Information using Machine Learning. *American Journal of Electronics & Communication, 1 (4), 16-21*

Dinh, T., Chakraborty, G. and McGaugh, M. (2020). Exploring Online Drug Reviews using Text Analytics, Sentiment Analysis and Data Mining Models. *SAS 2020 Global Forum, 4809*

Gräßer, F., Kallumadi, S., Malberg, H. and Zaunseder, S. (2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. *In Proceedings of the 2018 International Conference on Digital Health (DH '18). ACM, New York, NY, USA*

Islam, N., Haque, R., Pareek, P. K., Islam, M., Sajeeb, I. H. and Ratul, M. H. (2023). Deep Learning for Multi-Labeled Cyberbully Detection: Enhancing Online Safety. *In proceedings of 2023 International Conference on Data Science and Network Security (ICDSNS)*

Joshi, S. and Abdelfattah, E. (2021). Multi-class Text Classification Using Machine Learning Models for Online Drug Reviews. Retrieved from https://www.researchgate.net/publication/352621618 accessed on 13th February, 2024

Khanam, Z , Alwasel, B. N, Sirafi, H, and Rashid, M. (2021)" Fake News Detection Using Machine Learning Approaches" *IOP Conf. Series: Materials Science and Engineering 1099 012040 DOI:10.1088/1757-899X/1099/1/012040*

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3*

Mehta, D. K., Patel, M. B., Dangi, A., Patwa, N., Patel, Z., Jain, R., Shah, P. D. and Suthar, B. R. (2024). Exploring the efficacy of natural language processing and supervised learning in the classification of fake news articles. *Advanced of Robotics Technology, 2(1)*

Padalko, H., Chomko, V., Yakovlev, S. and Chumachenko, D (2023). Ensemble machine learning approach for fake news classification. *Intelligent information technologies, 4(108), 5-18*

*Sh*arma, S., Saran, Shankar M, and Patil (2020) "Fake News Detection using Machine learning algorithm: *International Journal of creative research thought (IJCRT), 8(6).*

Soladoye, A. A. (2023). Decision support system for prediction of stroke using Recurrent Neural Networks with Gated Recurrent Units. *Master Thesis, Department of Computer Engineering, Federal University, Oye-Ekiti. Nigeria*

Uddin, M. N., Hafiz, M. F., Hossain, S. and Mominul-Islam, S. M. (2022). Drug Sentiment Analysis using Machine Learning Classifiers. *(IJACSA) International Journal of Advanced Computer Science and Applications, 13(1), 92-100*

Vijayaraghavan, S. and Bas, D. (2020). Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms. *arXiv:2003.11643v1 [cs.CL]*