iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VIII **Month of publication:** August 2020

DOI: https://doi.org/10.22214/ijraset.2020.31141

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

Medical Diagnosis using Verbal Description to Classify the Category of the Ailment

Dr. Guru R[1], Nithish Bhat[2], Koushik J[3], Shreyas K[4]
[1]*Asst. Prof, Computer Science Department*
[2, 3, 4]*Students, JSS Science and Technology University, Mysore -570006*

Abstract: ***A huge amount of unstructured text data that contains valuable information over the online. This text is more volatile and changing rapidly making hard to process, read, extract and process. Natural language Processing (NLP) may be a subfield of linguistics which is worried with interactions with text data accustomed extract the key information. NLP helps to beat the challenges which is related to text data vulnerability. There are numerous publicly available unstructured text data. Among those we chose medical text data which is comprised of description of patients within the text and voice format. Basically, doctors follow an explicit procedure of collecting data of assorted symptoms or anomalies that a patient experiences so as to research true and treat the condition. Often the patient has vague memory of what and when these anomalies have occurred. we aim to form accurate records of those ailments together with when it occurred to assist doctors better understand the circumstance in as little time as possible. We used different machine and deep learning model to extract the summary from the statement text data to classify the ailments.***

Keywords: Machine Learning, Deep Learning, Natural Language Processing, Linguistic Analysis, Medical Text Classification.

I. INTRODUCTION

Internet Technology has been developing rapidly over the years and because of this there's an enormous amount text data being deployed online. Majorly these text data are classified into Structured, Semi-Structured and Unstructured text data. Unstructured text data has the bulk shares. So, this makes automatic text summarization and classification task extremely important. this is often achieved by tongue Processing which might understand the human language because it is spoken. It mainly deals with extracting, processing, classifying the text data. Traditionally to summarize a text data we've various approach like bag-of-words, Support vector machine, Naïve Bayes Classifier, Neural networks. Medical industry is flourishing nowadays adopting newest technology over the amount. there's significant amount of medical data available within the internet. These data possess some valuable information which is employed to research the varied diseases and symptoms.

Few of the related studies are focused on classifying the disease supported binary class problems. On the contrary, majority of the study focused on handling medical text on a sentence level. for instance, social media posts, tweets and question and answer. These were only limited to only classification. But summarizing the text data provides a more generified approach and applies to vast number of vocabularies. So, during this approach we've incorporated various deep learning and machine learning approaches to urge text summary from the information instead of classifying into limited categories.

In this paper, we aim at addressing some critical issues raised using machine learning algorithms for diagnosing and prediction. We start with examining the notion of interpret ability and the way it's associated with machine learning. Then, we provide a brief overview of the state of the art in medical AI. Against this background, we imply what we consider two crucial issues: the primary issue is that epistemic opacity is at odds with a typical desire of understanding and potentially undermines information rights. The second (related) issue concerns the assignment of responsibility in cases of failure. Subsequently, we elaborate these issues very well. Thereafter, we recommend that explainable AI might help to beat a number of the issues. Finally, we glance at several the implications for practice normally and for ailment classification.

We trained our summary model supported the medical dataset which consists of the 000 description of the patients. It consists of 6662 of the descriptions. We've got 24 classified diseases which maps to numerous descriptions.

To use this unstructured information in our summary model we used various machine learning and deep learning model. To feed these data to the model the information must be preprocessed. These preprocessing steps involves numerous steps like text sanitization, stop words/punctuation removal, sentence splitting, POS tagging, word tokenization, word lemmatization, and Named Entity Recognition (NER). After the preprocessing is completed the most step is to extract the useful features while discarding irrelevant and redundant features.

After this process we feed the features to coach the model to extract the summary and classify the ailment accordingly. We used various model to coach like statistical regression, Logistic regression, Support Vector Machine, Naïve Bayes Classifier and fast.ai classifier. We did a multimodal classification and took the simplest and efficient model to predict the result. during this case Logistic Regression achieved the very best efficiency with a accuracy of 99.78%.

II. RELATED WORK

In the paper named [1]"English Conversational Telephone Speech Recognition by Humans and Ma- chines" by George Saon, Gakuto Kurata, Tom Sercu Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, presented a group of acoustic and language modelling improvements to English speech with help of LSTM bidirectional models and it's also exploited complementarity between recurrent and convolutional architectures by adding word and character-based LSTM LMs and a convolutional WaveNet LM. Feature selection methods are classified in three types: filter, wrapper and hybrid approaches. Filter methods apply an independent test with none learning algorithm, while a pre-determined learning algorithm is employed by wrapper method. This paper helped us to a way to process the text data and to extract the most features from the conversational speeches. they need increased the classification accuracy compared to the present system in less time. The registrar general of India shows that the coronary cardiovascular disease could be a major explanation for death in our India which causes about 30 percent death in rural areas.

"Unstructured Medical Text Classification Using Linguistic Analysis: A Super- vised Deep Learning Approach" by Ahmad Al-Doulat, Islam Obaidat, and Min-woo Lee [2]. This paper presented the varied NLP processing methods mainly for unstructured text data and investigated the classification of online medical articles using linguistic (Semantic and Statistical) analysis. It showed that incorporating domain-specific terms and keywords can effectively improve the classification accuracy of the machine learning models when it involves specific domains. They Proposed the system which used the genetic search to prune redundant and irrelevant attributes also to seek out attributes which are important for classification. Least ranked attributes are removed, and classification is finished betting on high ranked attributes.

Also, for providing a recommendation system for the medication of the classified ailments we referred paper named "An Intelligent Medicine Recommender System Framework" by Youjun Bao and Xiaohong Jiang" [3]. This paper devises a universal medicine recommender system framework that applies data processing technologies to the diagnosing, which consists of database system module, data preparation module, recommendation model module, model evaluation model, and data visualization. Also, KNN based classifier accuracy depends on the worth of K and kind of distance metric. They used ten sample medical datasets. Results suggest the SVM model because the most desirable classification algorithm for developing a Recommendation framework. The research wasn't only to spot a classification algorithm that has been performing best altogether medical datasets but also reliability of the model. Hence managing clinical data, discovering patients' interactions, and integrating the various data sources were the most challenges. So, to beat these challenges they utilized the neural network model to form predictions on medical data.

III.METHODOLOGY

To begin we explore the contents of the dataset obtained from an opensource website appen[4]. The following subsections contains all the steps followed to clean the data and to train the model and obtain a multi class classification model with a high accuracy compared to existing ones. Most importantly the data is available in .wav audio format, it needs to be converted into text data.

A. Data Collection

To train the proposed model, we collected the dataset from an open source website called appen[4]. The dataset consists of 6662 phrases and corresponding prompt. The phrase involved contributors speaking and recording text phrases to describe symptoms and the prompt are the actual ailments that they are classified into, labels in other words. The data is available in excel format, but will need some pre-processing to be viable to train machine learning models. There are 25 labels, hence 25 classes that the phrases (sentences) can be classified into.

	file_name	phrase	prompt	overall_quality_of_the_audio	speaker_id
118	1249120_43788827_53247832.wav	Every time I take a deep breath I start coughing	Cough	3.33	43788827

Fig.1 Sample data

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

B. Data Overview

The dataset consists of 25 classes. These classes consist of common ailments that people could face. The classes are Acne, Back pain, Body feels weak, Cough, Back pain, Ear ache, Emotional pain, Feeling cold, Hair falling out, Hard to breath, Head ache, Head hurts, Infected wound, Injury from sports, Internal pain, Joint pain, Knee pain, Muscle pain, Neck pain, Open wound, Shoulder pain, Skin issue, Stomach ache, Feeling dizzy, Foot ache. The number of these prompts are as shown in fig.1. clearly the different labels are not equal in number which could affect the classification model in terms of bias.

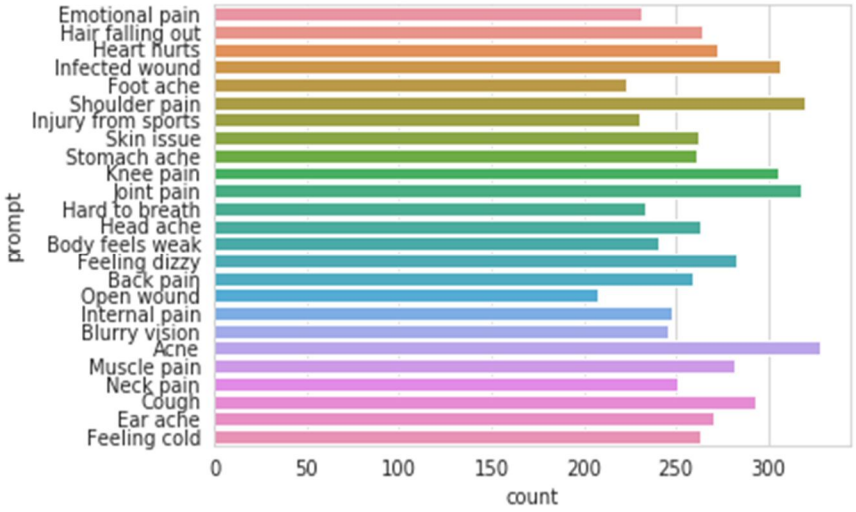


Fig.2 Data overview

C. Data Preprocessing

In this step, we preprocess the data to clean up our dataset. The preprocessing phase includes the following steps:

1) Stop Words Removal: to filter out useless words. These stop words are commonly used words (such as ” the”, ”a”, ”an”, ”in” etc.). For this step, we used the predefined list of stop words by NLTK.

2) Punctuation Removal: we use NLTK word tokenizer to pick out sequences of alphanumeric characters as tokens and drop everything else (punctuation).

D. Feature extraction form dataset using bag of words:

The most intuitive way to do so is to use a bags of words representation:

1) Assign a fixed integer id (a number) to each word occurring in any document of the training set (for instance by building a dictionary from words to integer indices).

2) For each document #X, count the number of occurrences of each word w and store it in E [i, j] as the value of feature #Y where Y is the index of word w in the dictionary.

E. Algorithms considered for training the model:

1) Logistic regression

2) Naïve Bayesian classifier

3) Neural network using fast ai classifier (pre trained model)

4) Support vector machine

F. Training the Model

In this step we define the model we are working with. The model we choose to train is a linear regression model. To compare the effectiveness of our approach we compare it to different machine learning models like the support vector machine, neural networks and Naive Bayes classifier. All the algorithms we use are multi class, which means that each sentence need to be classified into one among multiple classes. We use the library functions to implement these models, this gives us control and flexibility over the various parameters the algorithms have to offer in order to train the model.

The important point here is that the neural networks is capable of learning large amounts of data, but not necessarily with high accuracy. A simple algorithm such as logistic regression gives higher accuracy than the neural networks in the fraction of time that is required to train a pre-trained the neural network.

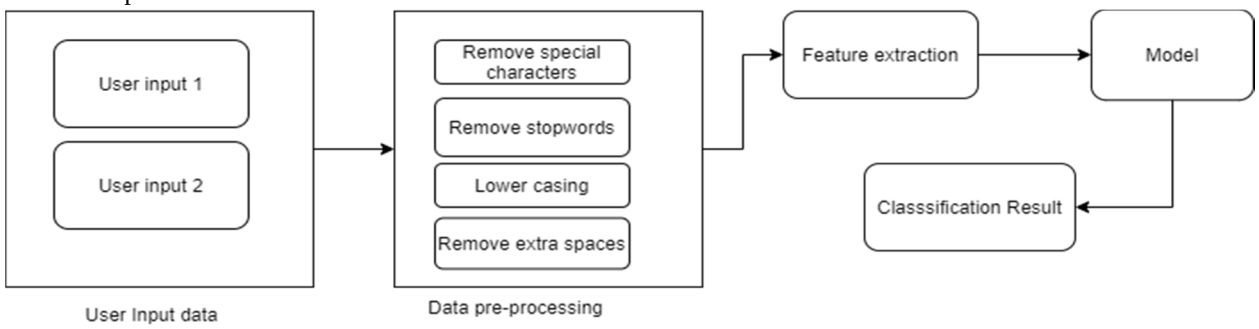


Fig.3 System design

G. General Steps Used for Training the Model

1) Read the dataset into a variable x using the Pandas library.

2) Preprocess the data using NLTK libraries, cleaning the text from stop words and.

3) Extract features from the dataset using bag of words and convert occurrence to frequencies.

4) Split the dataset into training and test dataset at 70 percent and 30 percent respectively.

x_train, x_test, y_train, y_test = train_test_split(Sentences, labels, test_size=xxx, random_state = xxx)

sentences: they are phrases which needs to be classified.

labels: they are the target classes.

test_size: used for splitting the data into test and training datasets.

random_state: parameter is used for initializing the internal random number generator, which will decide the splitting of data into train and test indices in your case.

5) In order to make the vectorizer => transformer => classifier easier to work with, pipeline class that behaves like a compound classifier:model = Pipeline ([("transform1", transformer_1), ("transform2", transformer_2), ("estimate", estimator)]).Transformers are classes that implement both fit () and transform (). You might be familiar with some of the preprocessing tools, like TfidfVectorizer and Binarizer. If you look at the docs for these preprocessing tools, you'll see that they implement both of these methods. Estimators are classes that implement both fit () and predict (). You'll find that many of the classifiers and regression models implement both these methods, and as such you can readily test many different models. Estimators can be assigned as Logistic regression, naïve Bayesian or support vector machine which results in the respective classifiers. Additional arguments include max iterations, jobs etc**.**

6) Use model.fit **(**x_train, y_train) function to begin the training with the split dataset which consists of 70 percent of the original datasetWhere

x_train: is an array of phrases that is used for training the model.

y_train: is an array of prompt (labels or classes) that is used for training the model.

7) Use classification_report(y_test, y_pred, target_names) function to get the accuracy ,precision and f1 score of the model on the test dataset. Where

y_test: 1 dimensional array consisting of test labels corresponding to the test phrases

y_pred: 1 dimensional array consisting of predictions or classifications that the models made. Target_names: also, a 1-dimensional array consisting of all the label names (class names).

8) The above parameters can also be used to plot graphs for further analysis.

H. Hardware Requirement

1) *Processor:* Quad core Intel Core i7 Skylake or higher (Dual core is manageable).

2) *RAM:* 6GB of RAM or higher.

3) *Storage:* Minimum 500 MB HDD (SDD is preferable for better performance).

4) *GPU:* Nvidia 9x or 10x series (preferable to use graphic card that supports CUDA toolkit).

I. Software Requirement

1) Operating System: Windows / Linux / MacOS

2) Python 3+: Programming language

3) NumPy 1.18.2: NumPy is the fundamental package for scientific computing with python

IV. RESULTS AND ANALYSIS

Since the dataset is not very large the machine learning algorithms did not take much time all for training. The testing data used is 30 percent of the data samples obtained from the dataset. The result of the testing is as shown in the figure.3. Logistic regression has the highest accuracy among other algorithms and neural networks has the least accuracy relative to other algorithms. The neural network also takes ten times more time to train a model. The naïve Bayesian classifier and the support vector machine have a respectable accuracy.

This has a very simple explanation. Though the neural network is more evolved, complex and capable of storing large number of features, it is not an effective method to obtain models for small dataset. The dataset consists of only 6662 data samples and is enough for training a highly accurate logistic regression model.

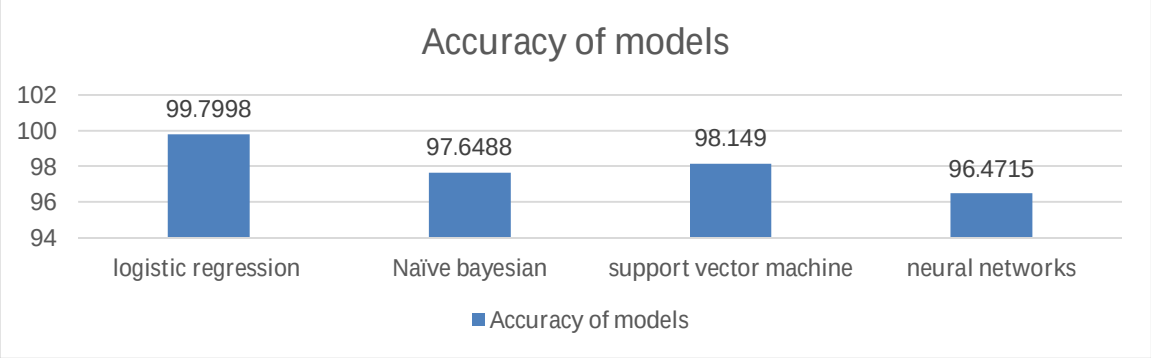


Fig.4 Final Testing Result

V. CONCLUSION

Unstructured Text Classification has attracted a serious attention within the present world. It's an important step in Natural Language processing for further analysis on unstructured text data. together with NLP various machine and deep learning models have provided a serious contribution within the medical field which helping in decreasing the fatality rate. With the assistance of Logistic regression text summary model, we help the people to predict the fundamental ailment for his or her symptoms. This model acts as a primary aid precaution for the people to self-diagnose their ailment and to require precautionary measures. This surely isn't a alternative for the medical professionals. This acts as a bridge between doctors and patients in analyzing the symptoms, to keep up record and diagnose efficiently. Lastly, we built an internet application using this algorithm to supply user with the top product. the net application is developed using HTML, CSS, JavaScript and used flask as middleware to fetch the records which are stored in MySQL database. This website is hosted on the Pythonanywhere.com platform and URL is http://red1998.pythonanywhere.com.

REFERENCES

[1] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, Phil Hall" English Conversational Telephone Speech Recognition by Humans and Machines".

[2] Ahmad Al-Doulat, Islam Obaidat, and Minwoo Lee" Unstructured Medical Text Classification Using Linguistic Analysis: A Supervised Deep Learning Approach".S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.

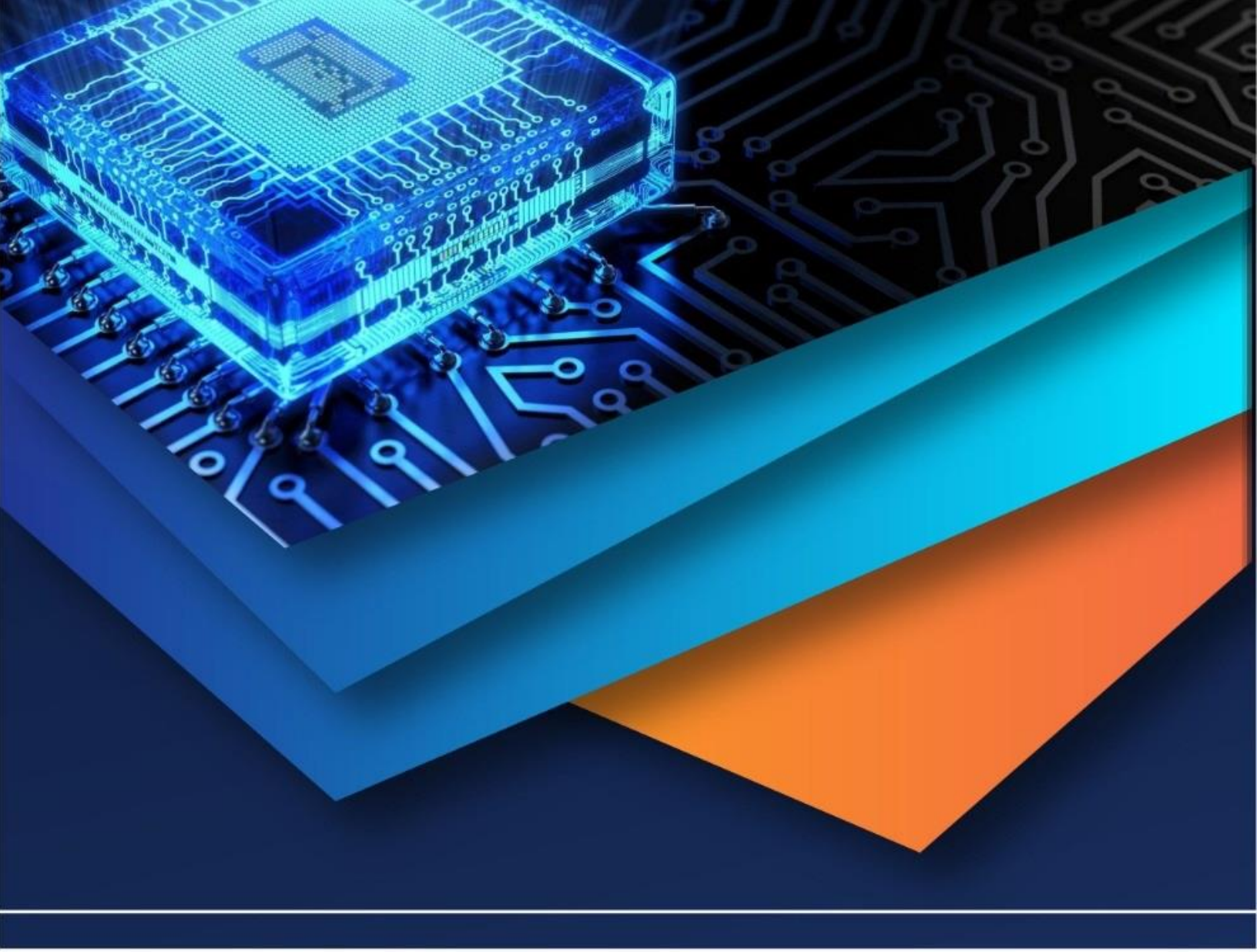
[3] Bao, Y., Jiang, X. (2016). "An intelligent medicine recommender system framework." 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA). doi:10.1109/iciea.2016.7603801.

[4] Appen datasets "Medical Speech, Transcription, and Intent (English)". Available: https://appen.com/datasets/audio-recording-and-transcription-for-medical-scenarios/.

[5] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," Machine learning, vol. 39, no. 2-3, pp. 103–134, 2000.

[6] S. Ananiadou, D. B. Kell, and J.-i. Tsujii, "Text mining and its potential applications in systems biology," Trends in biotechnology, vol. 24, no. 12, pp. 571–579, 2006.

[7] N. Jagannatha and H. Yu, "Structured prediction models for rnn based sequence labeling in clinical text," in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2016, p. 856, NIH Public Access, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY