Sentimental Analysis of Medicine Reviews using the Machine Learning Techniques

Naveen R[1], Anjan Babu G[2]

Dept of Computer Science, SV University, Tirupati

***Abstract*— *With the enormous growth of Internet, more users have engaged in health communities such as medical forums to gather health-related information, to share experiences about drugs, treatments, diagnosis or to interact with other users with similar condition in communities. Monitoring social media platforms has recently fascinated medical natural language processing researchers to detect various medical abnormalities such as adverse drug reaction. In this paper, we present a benchmark setup for analyzing the sentiment with respect to users' medical condition considering the information, available in twitter. To this end, we have crawled the medical forum website 'patient.info' with opinions about medical condition self narrated by the users. We constrained ourselves to some of the popular domains such as depression, anxiety, asthma, and allergy. The focus is given on the identification of multiple forms of medical sentiments which can be inferred from users' medical condition, treatment, and medication. Thereafter, a deep Convolutional Neural Network (CNN) based medical sentiment analysis system is developed for the purpose of evaluation. The resources are made available to the community through LRE map for further research.*

I. INTRODUCTION

Attention towards sentiment analysis has been flourishing over the last two decades because of the immense popularity of social media. The phenomenal rise in blogging trend is observed in health communities such as medical forums which are swamped by millions of users (many of whom are patients) seeking for health-related information, sharing medical problems or experiences and opting for in-formational support or opinions from the other users (patients, health-professional or doctors).

Medical sentiment analysis has its major applications in assessing the clinical records and in providing an automated decision support system for health professional. According to the study conducted by the Pew Internet & American Life Project1, almost 80 percent of Internet users in US have explored health-related topic online. More often, people look for the information about specific medical problem (63%) over the internet. Nearly 47% of the users search for the medical treatment or procedure in the internet. With such a tremendous amount of freely available medical texts in the web, it is necessary to harness the crucial and important information. Analyzing these texts by capturing the sentiments is helpful because opinions are central to almost all human activities and are key influencer of our behaviors. Although, several techniques exist to capture sentiments in general domains, the sentiments expressed in medical narratives have not been well analyzed and exploited in the required measure as yet.

Sentiment analysis is also widely known as opinion mining within NLP which tries to identify and extract the opinion within a text. Machine Learning algorithms are broadly divided into supervised and unsupervised learning methods. The supervised models require a training dataset or also known as labels which is used to derive relationship between the features to predict an output, whereas an unsupervised model don't use these labels instead cluster the data into common groups. Another category of learning is the semis supervised where unsupervised are used to develop labels for supervised machine learning.

The models fall into two main categories the conventional machine learning model and the deep learning models. With recent developments is various deep learning models the ability to apply deep learning to NLP and analyzing text has increased drastically. In recent applications the deep learning models with methods where words are projected as vectors has led to impressive results.

Taking these techniques into account this project aims at applying 2 conventional methods such as the CNN (Convolutional Neural Networks) in dense layer.

II. LITERATURE REVIEW

Medical Imaging using Machine Learning and Deep Learning Algorithms: A Review

Author: Jahanzaib Latif, Chuangbai Xiao, Azhar Imran, Shanshan Tu

Machine and deep learning algorithms are rapidly growing in dynamic research of medical imaging. Currently, substantial efforts are developed for the enrichment of medical imaging applications using these algorithms to diagnose the errors in disease diagnostic systems which may result in extremely ambiguous medical treatments. Machine and deep learning algorithms are important ways in medical imaging to predict the symptoms of early disease. Deep learning techniques, in specific convolutional networks, have promptly developed a methodology of special for investigating medical images. It uses the supervised or unsupervised algorithms using some specific standard dataset to indicate the predictions. We survey image classification, object detection, pattern recognition, reasoning etc. concepts in medical imaging. These are used to improve the accuracy by extracting the meaningful patterns for the specific disease in medical imaging. These ways also indorse the decision-making procedure. The major aim of this survey is to highlight the machine learning and deep learning techniques used in medical images. We intended to provide an outline for researchers to know the existing techniques carried out for medical imaging, highlight the advantages and drawbacks of these algorithms, and to discuss the future directions. For the study of multi-dimensional medical data, machine and deep learning provide a commendable technique for creation of classification and automatic decision making. This paper provides a survey of medical imaging in the machine and deep learning methods to analyze distinctive diseases. It carries consideration concerning the suite of these algorithms which can be used for the investigation of diseases and automatic decision-making.

Review of Medical Decision Support and Machine-Learning Methods

Author: Abdullah Awaysheh, Jeffrey Wilcke, François Elvinger

Machine-learning methods can assist with the medical decision-making processes at the both the clinical and diagnostic levels. In this article, we first review historical milestones and specific applications of computer-based medical decision support tools in both veterinary and human medicine. Next, we take a mechanistic look at 3 archetypal networks—commonly learning algorithms—naive Bayes, decision trees, and neural inner workings used to power these medical decision support tools. Last, we focus our discussion on the data sets used to train these algorithms and examine methods for validation, data representation, transformation, and feature selection. From this review, the reader should gain some appreciation for how these decision support tools have and can be used in medicine along with insight on their inner workings.

A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques

Author: Ambika Choudhury, Deepak Gupta

While designing medical diagnosis software, disease prediction is said to be one of the captious tasks. The techniques of machine learning have been successfully employed in assorted applications including medical diagnosis. By developing classifier system, machine learning algorithm may immensely help to solve the health-related issues which can assist the physicians to predict and diagnose diseases at an early stage. We can ameliorate the speed, performance, reliability, and accuracy of diagnosing on the current system for a specific disease by using the machine learning classification algorithms. This paper mainly targets the review of diabetes disease detection using the techniques of machine learning. Further, PIMA Indian Diabetic dataset is employed in machine learning techniques like artificial neural networks, decision tree, random forest, naïve Bayes, k-nearest neighbors, support vector machines, and logistic regression and discussed the results with their pros and cons.

Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review

Author: Indranil Balki,Afsaneh Amirabadi, Jacob Levman

The required training sample size for a particular machine learning (ML) model applied to medical imaging data is often unknown. The purpose of this study was to provide a descriptive review of current sample-size determination methodologies in ML applied to medical imaging and to propose recommendations for future work in the field. There were only 4 studies that discussed sample-size determination methodologies, and 18 that tested the effect of sample size on model performance as part of an exploratory analysis. The observed methods could be categorized as pre hoc model-based approaches, which relied on features of the algorithm, or post hoc curve-fitting approaches requiring empirical testing to model and extrapolate

algorithm performance as a function of sample size. Between studies, we observed great variability in performance testing procedures used for curve-fitting, model assessment methods, and reporting of confidence in sample sizes.

Medical Diagnosis Using Machine Learning: A Statistical Review

Author: Kaustubh Arun Bhavsar , Jimmy Singla , Yasser D. Al-Otaibi , Oh-Young Song, Yousaf Bin Zikria and Ali Kashif Bashir

Decision making in case of medical diagnosis is a complicated process. A large number of overlapping structures and cases, and distractions, tiredness, and limitations with the human visual system can lead to inappropriate diagnosis. Machine learning (ML) methods have been employed to assist clinicians in overcoming these limitations and in making informed and correct decisions in disease diagnosis. Many academic papers involving the use of machine learning for disease diagnosis have been increasingly getting published. Hence, to determine the use of ML to improve the diagnosis in varied medical disciplines, a systematic review is conducted in this study. To carry out the review, six different databases are selected. Inclusion and exclusion criteria are employed to limit the research. Further, the eligible articles are classified depending on publication year, authors, type of articles, research objective, inputs and outputs, problem and research gaps, and findings and results. Then the selected articles are analyzed to show the impact of ML methods in improving the disease diagnosis. The findings of this study show the most used ML methods and the most common diseases that are focused on by researchers. It also shows the increase in use of machine learning for disease diagnosis over the years. These results will help in focusing on those areas which are neglected and also to determine various ways in which ML methods could be employed to achieve desirable result

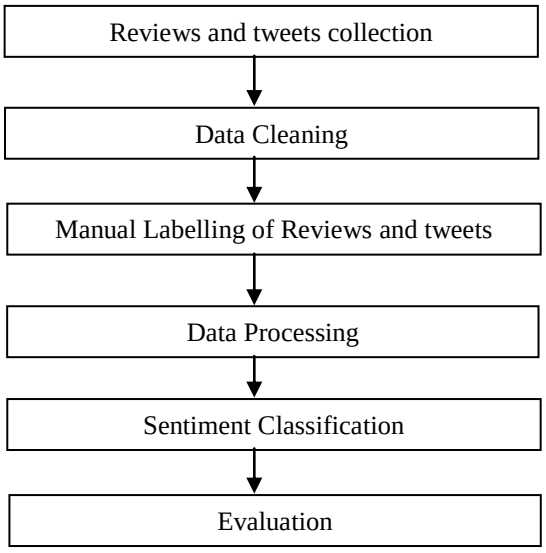
Problems in Earlier System

- The researchers that use pattern recognition of these data mining methods help in predicting models based on the medical reviews.
- The experiments that were carried out using these classification-based algorithms such as Dense layer, Sequential, SoftMax.
- These results have proven to be that of dense layer technique that have performed better than the others when utilized by the techniques

Disadvantages

- The existing system used different algorithm to predict the disease, but accuracy is low comparison of our model.
- Feature Extraction is very complex.
- Training and testing the model is used same algorithm, but we provide different method.

III. ARCHITECTURE



IV. PROPOSED WORK

- Our proposed system involves Dense Layer in Convolutional Neural Network (CNN) Algorithm in Deep Learning concept used to train the dataset.
- In **Dense Layer**, each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers.
- In Dense Layer uses features of all complexity levels. It tends to give more smooth decision boundaries.

Advantages

- Easy detection of the medical reviews with the concluded technique.
- Time consuming.
- Best accuracy Model helps in better treatment as early.
- Detection of best Model will quick the treatment which is life saving.

4.1 Dataset Collection

A dataset (or data set) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the dataset in question. It lists values for each of the variables, such as height and weight of an object. Each value is known as a datum.We have chosen to use a publicly-available Healthcare dataset which contains a relatively small number of inputs and cases. The data is arranged in such a way that will allow those trained in medical disciplines to easily draw parallels between familiar statistical and novel ML techniques. Additionally, the compact dataset enables short computational times on almost all modern computers**.**

4.2 Pre-Processing

The Keruspre-processing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the set, robust scalers or transformers are more appropriate. The behaviors of the different scalers, transformers, and normalizers on a dataset containing marginal outliers is highlighted in Compare the effect of different scalers on data with outliers.

4.2.1 Standardization, or Mean removal and Variance Scaling

Standardization of datasets is a **common requirement for many machine learning estimators** implemented in scikit-learn; they might behave badly if the individual features do not more or less look like standard normally distributed data: Gaussian with **zero mean and unit variance.**

4.2.2 Scaling features to a range

In practice we often ignore the shape of the distribution and just transform the data to center it by removing the mean value of each feature, then scale it by dividing non-constant features by their standard deviation. For instance, many elements used in the objective function of a learning algorithm (such as the RBF kernel of Support Vector Machines or the l1 and l2 regularizers of linear models) assume that all features are centered around zero and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected. An alternative standardization is scaling features to lie between a given minimum and maximum value, often between zero and one, or so that the maximum absolute value of each feature is scaled to unit size. This can be achieved using MinMaxScaler or MaxAbsScaler, respectively. The motivation to use this scaling include robustness to very small standard deviations of features and preserving zero entries in sparse data. MaxAbsScaler works in a very similar fashion, but scales in a way that the training data lies within the range [-1,1] by dividing through the largest maximum value in each feature. It is meant for data that is already centered at zero or sparse data.

4.3 Normalization

Normalization is the process of **scaling individual samples to have unit norm**. This process can be useful if you plan to use a quadratic form such as the dot-product or any other kernel to quantify the similarity of any pair of samples.

This assumption is the base of the Vector Space Model often used in text classification and clustering contexts.

4.4 EDA (Exploratory Data/Visualized Analysis)

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modelling the data in order to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter plot and many more. It often takes much time to explore the data. Through the process of EDA, we can ask to define the problem statement or definition on our data set which is very important. So, in this tutorial, we will explore the data and make it ready for modelling.

4.4.1 Importing the required libraries for EDA

The libraries that are used in order to perform EDA (Exploratory data analysis).

4.4.2 Loading the data into the data frame.

Loading the data into the pandas' data frame is certainly one of the most important steps in EDA, as we can see that the value from the data set is comma-separated. One thing to remember in this step is that uploaded files will get deleted when this runtime is recycled.

4.4.3 Checking the types of data

The datatypes because sometimes the MSRP or the price of the car would be stored as a string or object, if in that case, we have to convert that string to the integer data only then we can plot the data via a graph.

4.4.4 Dropping irrelevant columns

This step is certainly needed in every EDA because sometimes there would be many columns that we never use in such cases dropping is the only solution. In this case, the columns such as Engine Fuel Type, Market Category, Vehicle style, Popularity, Number of doors, Vehicle Size doesn't make any sense to me so I just dropped for this instance**.**

4.4.5 Renaming the column

In this instance, most of the column names are very confusing to read, so I just tweaked their column names. This is a good approach it improves the readability of the data set.

4.4.6 Dropping the duplicate rows

This is often a handy thing to do because a huge data set as in this case contains more than 10, 000 rows often have some duplicate data which might be disturbing, so here I remove all the duplicate value from the data-set. For example, prior to removing I had 11914 rows of data but after removing the duplicates 10925 data meaning that I had 989 of duplicate data.

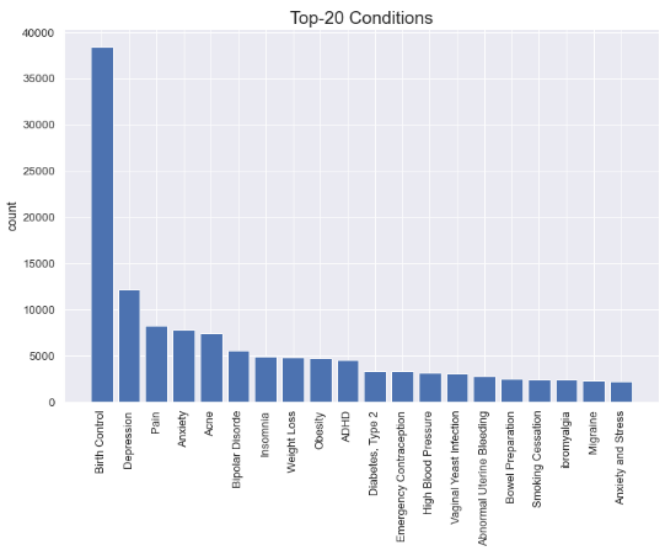
4.4.7 Dropping the missing or null values.

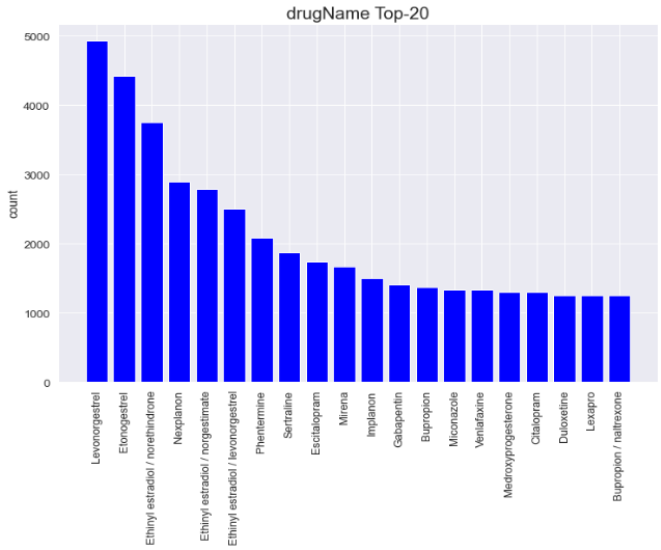
An outlier is a point or set of points that are different from other points. This is mostly similar to the previous step but in here all the missing values are detected and are dropped later. Now, this is not a good approach to do so, because many people just replace the missing values with the mean or the average of that column

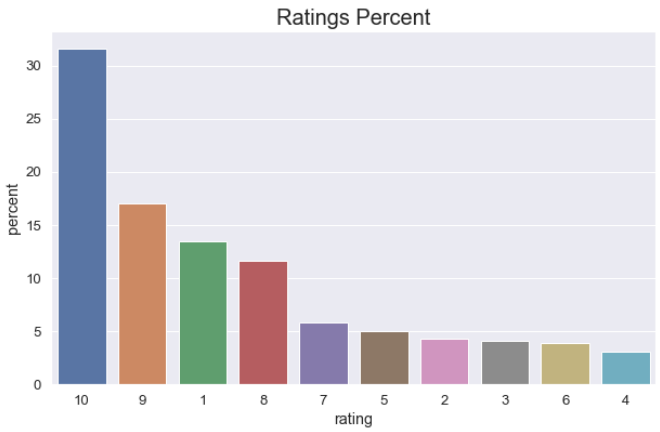
4.4.8 Detecting Outlier

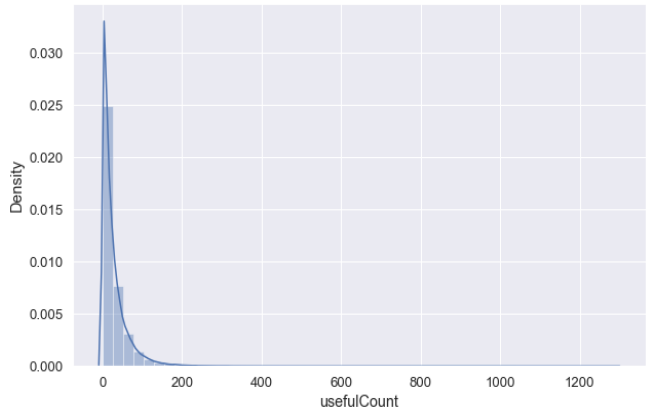
Sometimes they can be very high or very low. It's often a good idea to detect and remove the outliers. Because outliers are one of the primary reasons for resulting in a less accurate model. Hence, it's a good idea to remove them. The outlier detection and

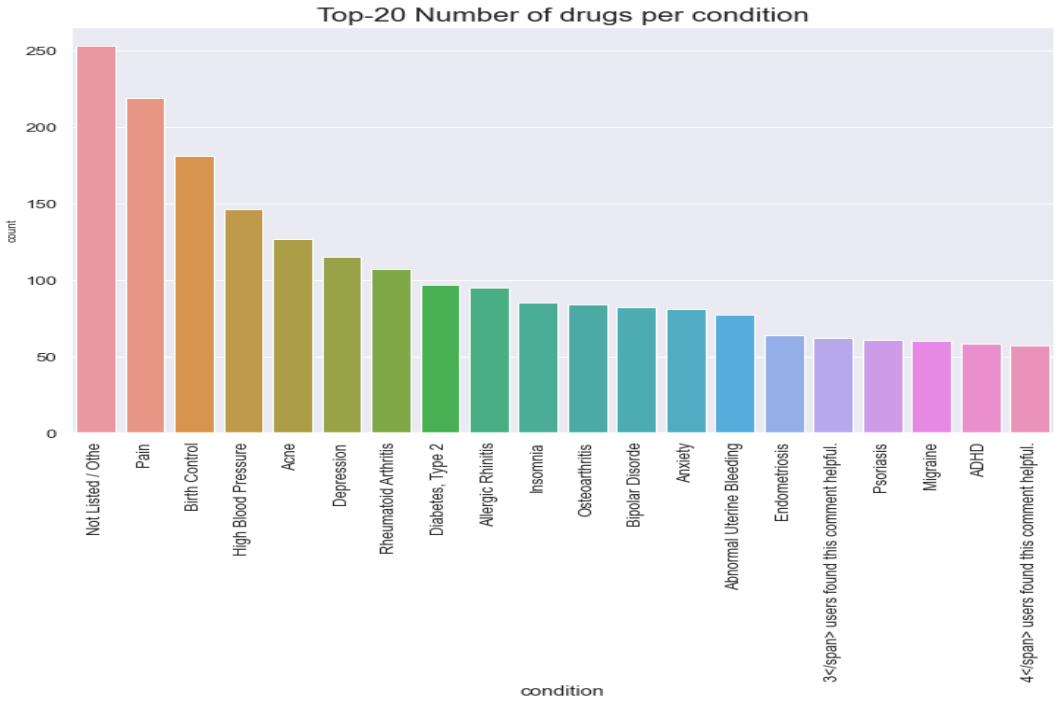
removing that I am going to perform is called IQR score technique. Often outliers can be seen with visualizations using a box plot.

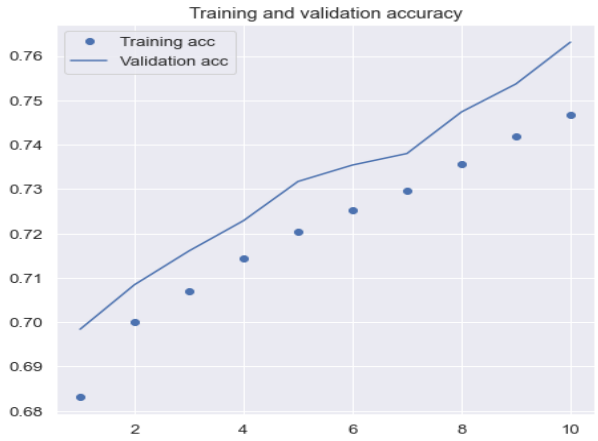


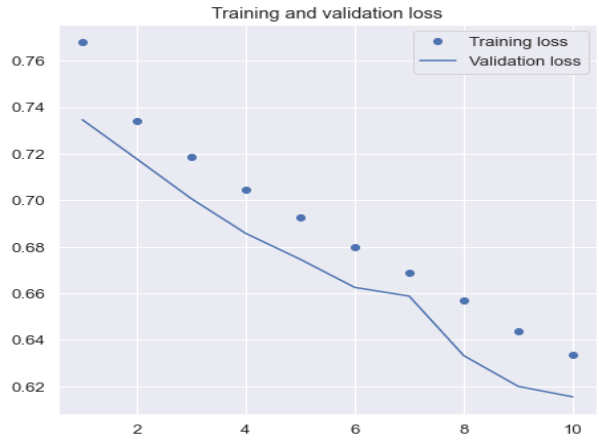












4.5 Implement The Model

The implementation model represents how a system (application, service, interface, etc.) works. It is often described with system diagrams and pseudocode to be later translated into real code. It is shaped by technical, organizations, and business constraints. Here we use sequential model, softmax. Sequence models the machine learning models that input or output sequences of data. Sequential data includes text streams, audio clips, video clips, time-series data and etc. The softmax function is often used in the final layer of a neural network-based classifier. Such networks are commonly trained under a log loss (or cross-entropy) regime, giving a non-linear variant of multinomial logistic regression. Softmax converts a real vector to a vector of categorical probabilities.

The elements of the output vector are in range (0, 1) and sum to 1.

Each vector is handled independently. The axis argument sets which axis of the input the function is applied along.

Softmax is often used as the activation for the last layer of a classification network because the result could be interpreted as a probability distribution.

The softmax of each vector x is computed as exp(x) / tf.reduce.sum(exp(x)).

The input values in are the log-odds of the resulting probability.

Arguments

- **x** : Input tensor.
- **axis**: Integer, axis along which the softmax normalization is applied.

Returns

Tensor, output of softmax transformation (all values are non-negative and sum to 1).

Raises

- **Value Error**: In case dim(x) == 1.

4.6 Sentiment Analysis

Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

4.7 Types of Sentiment Analysis

Sentiment analysis models focus on polarity (*positive, negative, neutral*) but also on feelings and emotions (*angry, happy, sad),* urgency *(urgent, not urgent) and* even intentions (*interested v. not interested*).

Depending on how you want to interpret customer feedback and queries, you can define and tailor your categories to meet your sentiment analysis needs. In the meantime, here are some of the most popular types of sentiment analysis:

4.8 Fine-grained Sentiment Analysis

If polarity precision is important to your business, you might consider expanding your polarity categories to include:

- Very positive
- Positive
- Neutral
- Negative
- Very negative

This is usually referred to as fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:

- Very Positive = 5 stars
- Very Negative = 1 star

4.9 Emotion detection

This type of sentiment analysis aims to detect emotions, like happiness, frustration, anger, sadness, and so on. Many emotion detection systems use lexicons (i.e., lists of words and the emotions they convey) or complex machine learning algorithms.

V. CONCLUSION AND FUTURE WORK

In this project a sentiment analysis was performed on the drug review dataset using various conventional and deep learning models to evaluate their performance. It was observed that the deep learning models performed better. The accuracy seemed to be enhanced when the vector representation was used in any method. Unequal class distribution was presumed to be a major issue in training and testing the dataset and thereby yielding lower accuracy. The presence of lesser dataset could have also caused to the significant reduction in the performance of the models than what was expected. Overall, the binary classification worked almost as what was expected and yielded better results in terms of accuracy when compared to the 3-class classification.

REFERENCES

[1] Adrover, Cosme, Todd Bodnar, Zhuojie Huang, AmalioTelenti, and Marcel Salathé. "Identifying adverse effects of HIV drug treatment and associated sentiments using Twitter." JMIR public health and surveillance 1, no. 2 (2015): e7.

[2] Sinnenberg, Lauren, Alison M. Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina M. Merchant. "Twitter as a tool for health research: a systematic review." American journal of public health 107, no. 1 (2017): e1-e8.

[3] Latif, Jahanzaib, Chuangbai Xiao, Azhar Imran, and Shanshan Tu. "Medical imaging using machine learning and deep learning algorithms: a review." In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1-5. IEEE, 2019.

[4] Choudhury, Ambika, and Deepak Gupta. "A survey on medical diagnosis of diabetes using machine learning techniques." In Recent developments in machine learning and data analytics, pp. 67-78. Springer, Singapore, 2019.

[5] Kaur, Prabhpreet, Gurvinder Singh, and Parminder Kaur. "A review of denoising medical images using machine learning approaches." Current medical imaging 14, no. 5 (2018): 675-685.

[6] Houssein, Essam H., Marwa M. Emam, Abdelmgeid A. Ali, and PonnuthuraiNagaratnamSuganthan. "Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review." Expert Systems with Applications (2020): 114161.

[7] Samant, Piyush, and Ravinder Agarwal. "Machine learning techniques for medical diagnosis of diabetes using iris images." Computer methods and programs in biomedicine 157 (2018): 121-128.

[8] Pillai, Rohan, ParitaOza, and Priyanka Sharma. "Review of machine learning techniques in health care." In Proceedings of ICRIC 2019, pp. 103-111. Springer, Cham, 2020.

[9] William, Wasswa, Andrew Ware, Annabella HabinkaBasaza-Ejiri, and JohnesObungoloch. "A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images." Computer methods and programs in biomedicine 164 (2018): 15-22.

[10] Islam, Md Aminul, and Nusrat Jahan. "Prediction of onset diabetes using machine learning techniques." International Journal of Computer Applications 975 (2017): 8887.

[11] Rahman, Atiqur, and Md Sharif Hossen. "Sentiment analysis on movie review data using machine learning approach." In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1-4. IEEE, 2019.