American Journal of Electronics & Communication

Shuvendu Das et. al., American Journal of Electronics & Communication, Vol. I (4), 16-21

Disease Prediction from Drug Information using Machine Learning

Shuvendu Das[1], Sainik Kumar Mahata[2], Abhishek Das[3], Koushik Deb[4]

[1, 2, 4]Institute of Engineering and Management, Kolkata, India

[3]Bengal College of Engineering, Durgapur, India

[1]getshuvendu97@gmail.com, [2]sainik.mahata@gmail.com

[3]abhishek93.das93@gmail.com, [4]koushik.deb@iemcal.com

Abstract

Drug reviews play a very important role in providing crucial medical care information for both healthcare professionals and consumers. Also, in the absence of an actual practicing healthcare professional, a consumer can look for an online review of drugs before making a purchase. But these reviews are generally unstructured in nature and often do not provide concise information on the disease/nature of the disease, the drugs are prescribed for. In this scenario, a learning model that can be trained to predict the disease/type of disease, when provided with a drug name and its corresponding review, becomes very important. To mitigate the above-mentioned issue, we present and compare various machine learning-based prediction models. Also, the performance of each of the models has been quantified using metrics such as precision, recall, F1-Score, and accuracy.

Keywords

Machine Learning, Deep Learning, Classification, Prediction, Drug Review

1. Introduction

Due to this pandemic situation, where the majority of the population has been confined to their homes and online procurement of goods for daily use is the new normal, reviews of such goods on the online space plays a huge role and acts as a metric by which users can select products without physically feeling and checking them. Moreover, in the case of online shopping for medicines, reviews become even more important as mining on less descriptive, unverified and bogus reviews may become lethal.

Also, due to the absence of certified medical practitioners, most of the users are relying heavily on self-medication. In such a scenario, it becomes very difficult for non-expert users to study reviews of a drug and identify the disease for which the drug is prescribed for. This situation arises as the medicine literature is extremely vast, it becomes very difficult for non-expert users to pin-point on the right medicine, after observing reviews and feedback for the same in the online forums.

Moreover, the reviews and feedback of drugs on online forums are unstructured, which may lead to non-uniform judgement and inability to classify the comments into meaningful insights.

To alleviate the above issues, we plan to develop machine learning models, which when trained using pre-processed and structured medicine reviews, will be able to uniformly predict the name of the disease, for which its usage is intended for. This will serve the objective of enabling users to safely intake the predicted drug with utmost benefit. This will also ensure that users are not bogged down by unexpected side-effects as a result of wrong ingestion.

For developing the learning models, we have used two approaches; one model uses the traditional machine learning algorithms and the second one uses the more recent state-of-the-art deep learning algorithms. When validated, both our models returned high efficiency for metrics such as precision, recall, accuracy and F1 score.

The rest of the paper is organized as follows. Section 2 mentions some of the related work done on this domain in recent years. Section 3 talks about the data source that we used to develop our models and the process of cleaning the data. Section 4 describes the methodology followed for developing the model and also discusses the results of the models. The paper ends with the concluding remarks in Section 5.

2. Related Work

Hu et. al. [1], designed an unsupervised anomaly detection system using a drug review dataset, where they used probability density estimation models to describe the distribution of the data over a number of key attributes and use the model to identify anomalies as points with low estimated probability. The results are validated against cases identified by healthcare domain experts. There was strong agreement between cases identified by the models and expert clinical assessment.

Gräßer et. al. [2], in their work performed multiple tasks over drug reviews with data obtained by crawling online pharmaceutical review sites. They performed sentiment analysis to predict the sentiments concerning overall satisfaction, side effects and effectiveness of user reviews on specific drugs. To meet the challenge of lacking annotated data they further investigated the transferability of trained classification models among domains, i.e., conditions, and data sources. With this work they showed that transfer learning approaches can be used to exploit similarities across domains and is a promising approach for cross-domain sentiment analysis.

Dinh et. al. [3], developed a data-mining model to evaluate the effectiveness and detect potential side effects from online customer reviews on specific prescription drugs. The study utilizes text parsing, text filtering, text topic, and text clustering within SAS Enterprise Miner for feature engineering and supervised learning algorithm for building multiple predictive models (logistic regression, decision tree, neural network, text rule builder) to identify the optimal model for reviews classification. The study's results show that the best predictive model for side effect classification is the text rule builder model with a validation average square error of 5.79% and a misclassification rate of 31.57%. Regarding effectiveness classification, the text rule builder model also works best with 5.10% validation average square error and 29.08% misclassification rate.

Yadav et. al. [4], we present a benchmark setup for analyzing the sentiment with respect to users' medical condition considering the information, available in social media in particular. To this end, we have crawled the medical forum website 'patient.info' with opinions about medical conditions 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 the LRE map for further research.

While most of the recent works on drug reviews focus on sentiment analysis, evaluating side effects and effectiveness of drugs and predicting names of drugs, this is the first work on predicting the disease name from drug reviews that will help users select the optimal drug based on the ailment that they are suffering from.

3. Dataset

For developing our machine learning models, we used the UCI ML Drug Review dataset[1] that had 2,15,063 number of reviews along with drug name and the disease name. Since the data was unstructured and had multiple rogue textual spans, we first needed to pre-process the data to make it suitable for training purposes. These steps included the removal of extra characters to clean the data. The extra characters that

[1] https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

were removed/cleaned included mentions, punctuations and URLs. Also, words from hashtags were extracted and extra spaces were contracted.

After cleaning, 53, 498 structured reviews, along with drug name and disease names were extracted. Out of these, 52, 498 such instances were selected as training data and 1,000 instances were selected as test data. As far as the labels were concerned, there were 648 unique labels that were to be predicted.

4. Methodology and Results

For developing the multi-label, machine learning classification models, we used two approaches. The first approach incorporated the traditional machine learning models such as Random Forest and Naïve Bayes algorithms. The second approach incorporated the more state-of-the-art sequential Neural Network algorithm, embodying Long-Short term memory [5] (LSTM) cells. Description of both these approaches are provided in the following subsections.

4.1 Machine Learning model

Since the training data consisted of the drug name and the reviews of the same drug that needed to be mapped to the disease name that was kept as the label, we first decided to concatenate both the drug name and the review. After concatenation, the extended input took the following structure, where the review followed the drug name and an "equal to" sign.

> "Mobic = Reduced my pain by 80% and lets me live a normal life again!"

Also, the labels were passed through a Label Encoder, which encodes target labels with a value between 0 and n_classes - 1, where n_classes in our case was 648.

Thereafter, TF-IDF Vectorizer[2] from the python package sklearn[3], which converts a collection of raw documents to a matrix of TF-IDF features, was used to vectorize the extended input.

Subsequently, the Random Forest algorithm, which is an ensemble learning model, where we can create many decision trees and predict based on the highest voting, was used to build the classification model. We took 100 estimators or trees and 1000 depths to make predictions. This model garnered an accuracy figure of 0.83 when

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[3] https://scikit-learn.org/stable/

tested using the test data. Similarly, F1 score reached 0.75 and precision, recall stopped at 0.89 and 0.72.

Also, the Multinomial Naïve Bayes algorithm was also used to build the same classification model. This model garnered an accuracy score of 0.75 when tested using the test data. Also, F1 score reached to 0.58 and precision and recall stood at 0.82 and 0.59 respectively.

It was noticed that scores of the Random Forest algorithm were better than Naïve Bayes algorithm.

4.2 Deep Learning model

For developing the deep learning model, we used a sequential feed-forward network built using LSTM cells. In this case, the input features, drug names and the reviews, were kept separate and not concatenated, as in the previous experiment. Also, the reviews were passed to through the pre-trained Google News Word Embedding[4] of size 300. This, in turn, produced word vectors of size 300 for the reviews.

Also, a straight-forward model, where no pre-trained embedding was used, was alsodeveloped.

The name of the drug was first passed through an embedding layer. The output of this layer was then concatenated to the word vector of the reviews (both Google News embedding and default embedding) and were passed through two LSTM layers. The output of the LSTM layer was then passed to a dense layer, which mapped the context vector to the respective labels.

The model, with the pre-trained embedding garnered F1 score of 0.80 and scores of 0.82 and 0.79 for precision and recall respectively, when tested using the test data.

Similarly, the model with the default embedding, garnered F1 score of 0.79, and precision and recall scores of 0.83 and 0.77 respectively, when tested using the testdata.

A comparative analysis of the classification scores is shown in Table 1.

[4] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.83	0.89	0.72	0.75
Naive Bayes	0.75	0.82	0.59	0.58
LSTM + Google News embedding	0.81	0.82	0.79	0.80
LSTM + default embedding	0.79	0.83	0.77	0.79

Table 1: Comparison of the classification metrics achieved by the developed models.

5. Conclusion

In the presented work, we have developed machine learning models that can be used to predict the name of the disease, given the drug name and the review of the drug. This will be especially helpful for users, who do not have specialized advice of medical practitioners. Also, the users, by giving names of medicine and certain reviews of the same as inputs, will be able to correctly predict the name of a disease. Thereafter, after successful matching of the observed disease and the predicted disease, the user will be able to correctly order medicines online.

We have worked with various machine learning techniques and have observed that the model, encompassing the Random Forest algorithm, gave us the best prediction results. In contrast, the deep learning models came close in terms of accuracy, but could not outperform the traditional machine learning methods, as more often than not, DL methods rely on huge amounts of data for learning patterns.

In the future, we would like to experiment with more dataset and specialized medical domain embeddings created using state-of-the-art embedding models such as BERTand RoBERTa.

References

[1] Hu, X., Gallagher, M., Loveday, W., Connor, J.P. and Wiles, J., 2015, February. Detecting anomalies in controlled drug prescription data using probabilistic models. In *Australasian Conference on Artificial Life and Computational Intelligence* (pp. 337-349). Springer, Cham.

[2] Gräßer, F., Kallumadi, S., Malberg, H. and Zaunseder, S., 2018, April. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 International Conference on Digital Health* (pp. 121-125).

[3] Dinh, T., Detecting Side Effects and Evaluating Effectiveness of Drugs from Customers' Online Reviews using Text Analytics and Data Mining Models.

[4] Yadav, S., Ekbal, A., Saha, S. and Bhattacharyya, P., 2018, May. Medical sentiment analysis using social media: towards building a patient assisted system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[5] Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*,*9*(8), pp.1735-1780.