Predicting the Efficacy Rate of Different Drugs based on Patient Reviews

Agnideep Mukherjee
Student Id: 22260149
agnideep.mukherjee2@mail.dcu.ie

Andrei Radoi
Student Id: 22263516
andrei.radoi2@mail.dcu.ie

Anjali Kumari
Student ID: 22261032
anjali.kumari2@mail.dcu.ie

Oana Diaconescu
Student Id: 22260715
oana.diaconescu2@mail.dcu.ie

***Abstract*—Accurately predicting drug efficacy can help ensure that the most effective drugs are prescribed to patients. This can lead to improved treatment outcomes and better patient care. While clinical trials provide critical information about the effectiveness of a drug, they are often expensive, time-consuming, and may not fully capture the experience of patients using the drug in real-world settings. On the other hand, patient-reported outcomes, such as reviews and ratings, can provide valuable insights into drug efficacy and help personalize treatment. In this paper, we explore the use of natural language processing (NLP) techniques to extract relevant information from patient reviews and ultimately predict drug efficacy. We apply the Knowledge Discovery in Database (KDD) methodology to process the data and then investigate the effectiveness of various data mining models: Random Forests, Naive Bayes and K-Nearest Neighbours. The models' performance is evaluated on a real-world dataset, "UCI ML Drug Review", obtained by crawling the pharmacology website Drugs.com. In our study we evaluated the models' performance on two distinct tasks: categorizing the sentiment of reviews as positive or negative and classifying drug rating scores on a scale of 1-10. Our findings indicate that Random Forests outperformed the other models in both cases, achieving a high accuracy rate of more than 90% for binary classification and 75% for multi-class classification.**

***Index Terms*—Data Mining, Random Forests, Naive Bayes, K-Nearest Neighbours, Sentiment analysis**

I. Introduction

Each year, the pharmaceutical industry allocates considerable financial and material resources towards testing the safety and efficacy of new drugs. Extensive pre-clinical and clinical trials are conducted to evaluate potential side effects and determine the success rate of new products before their release on the market. Nevertheless, findings are often generalized based on a restricted target group, which may not necessarily be applicable to the broader population due to variables such as genetic variation, medical preconditions, or environmental factors. Accurately predicting drug efficacy is crucial for improving the drug development process and reducing costs. Therefore, the industry is actively seeking ways to improve current clinical evaluation practices.

Machine learning models have demonstrated great potential in predicting the efficacy of new drugs. These systems are capable of processing vast amounts of information and generating personalized medical recommendations by taking into consideration a variety of different factors. For instance, genomic analysis is being used to analyze a patient's DNA sequences and identify connections between an individual's genetic profile and their clinical response to specific drugs. Image analysis is a technique that can be used to examine visual features of medical images, such as MRIs or X-rays, to assess the impact of a treatment on the target organ or tissue. One major challenge with these approaches is that they usually rely on structured data, generally extracted from official medical records and formal reports. Unfortunately, most of the times there is no centralized database for this information, and even when such databases exist, patient records are confidential and protected by strict privacy and regulatory concerns.

To address these limitations, recent research has directed its focus towards the use of unstructured data for predicting drug efficacy. Particularly, natural language processing (NLP) techniques such as sentiment analysis are being used to extract subjective information from textual data. Sentiment analysis can be used to predict the sentiment of patients' reviews, comments, or feedback related to their experiences with a particular drug. This data can provide a better indication of a drug's efficacy compared to the information gathered from clinical trials or laboratory studies. This is because the latter methods may not accurately represent the drug's performance in diverse populations or real-life situations. Moreover, using machine learning algorithms trained on patients' descriptions can potentially capture more detailed information about a drug's effectiveness, as patients tend to provide more nuanced descriptions of their experiences with the drug compared to traditional metrics such as success rates or adverse events.

In this paper we will be applying sentiment analysis techniques to predict the efficacy of different drugs based on patients' feedback gathered by crawling online pharmaceutical review sites. Each drug is evaluated through a textual review and a ten point rating corresponding to its effectiveness. Our goal is to predict this rating based on a given patient description, which eventually allows us to transform unstructured data into quantitative information which can be used to analyze the effectiveness of different drugs. We will be comparing three different machine learning models - Naive Bayes, Random Forests and K-Nearest Neighbours - and evaluating their performance in predicting the correct scores.

The structure of this paper can be outlined as follows: Section II provides a summary of previous research studies that are relevant to the present work, Section III gives a concise

description of the dataset and the procedures implemented to train the models, Section IV details the outcomes and describes the methodology used for evaluation, and finally, Section V concludes the research findings.

II. Related Work

The earliest work on predicting drug efficacy from textual data was based on linguistic models. These type of models rely on pre-built lexicons of words and phrases that can be associated with a specific sentiment or attribute and can be used to recognize features of a given text by matching its words with entries from the lexicon [10].

Asghar et. al. [3] propose a hybrid approach for creating a sentiment lexicon for health-related content. An initial set of words is extracted from online sources, filtered using Unified Modeling Language System (UMLS) concepts and matched with corresponding sentiments from the SentiWordNet (SWN) dictionary. The proposed lexicon demonstrates improved results for both manual and public datasets, but is limited by the generalized nature of the web repositories, which can potentially include irrelevant words. Liu et.al. [4] propose a new feature extraction method that combines a medical sentiment lexicon and position encoding to extract features from drug review data. Comparative experiments using supervised learning algorithms show that the proposed method outperforms conventional frequency and simple chronological vector representation techniques. However, the study is limited by the lack of public medical text datasets with ground truth values and requires further work to improve data balance and evaluate the proposed method with datasets generated from various data sources. Saad et.al. [5] propose a hybrid approach combining lexicon-based and learning-based techniques for sentiment analysis of drug reviews. The study evaluates the effectiveness of general-purpose sentiment lexicons (AFFIN, TextBlob, and VADER) for data annotation and various feature engineering techniques such as TF-IDF for feature extraction. The results show that TextBlob provides better annotation results, and the combination of learning-based and lexicon-based approaches outperforms their individual use, with an improvement of 3% in accuracy compared to previous approaches.

Another popular approach to drug efficacy prediction is aspect-based sentiment analysis, which associates a sentiment to each feature extracted from a text. Imani et.al. [8] present a new method for aspect-based sentiment analysis (ABSA) of patient opinions, classifying implicit and explicit aspects of a drug, such as effectiveness, side effects, or dosage. The proposed method uses distant supervision to automate the construction of a training set using sentences and phrases annotated as aspect classes in the drug domain. Han et. al. [7] propose a new dataset called SentiDrugs for aspect-level sentiment analysis in drug reviews. Their paper introduces a multi-task learning model based on double BiGRU, which uses attention mechanism to generate target-specific representations for sentences, and transfers knowledge from short text-level tasks to improve the performance of aspect-level sentiment classification. Sweidan et.al. [9] propose an aspect-based approach which uses a combination of a lexicalized ontology and XLNet models for feature extraction, as well as Bidirectional Long Short Term Memory (Bi-LSTM) networks for classification. This solution achieves an accuracy of 98% on data extracted from social media texts, but has the limitation of low performance rate for XLNet when a long context is required. Gräßer et.al [6] propose an approach that combines cross-domain and cross-data learning to improve the performance of aspect-based sentiment analysis. One important finding in this work is that domain-specific vocabulary has a great impact on the generalization capabilities of a machine learning model, as results show that in-domain testing outperforms all cross-domain setups. Moreover, cross-data evaluation shows unsatisfactory results when generalizing models trained on small datasets, which suggests that more powerful deep learning models are required for this case.

Lately, machine learning models have gained popularity for predicting drug efficacy. For instance, Marthin et.al. [11] explored the use of machine learning to predict overall drug performance using unstructured textual data from online product reviews. The authors tested various machine learning models on a drug review dataset, including CART, C5.0, GLM, MARS, SVM, and Random Forest, and found that the SVM model with linear kernel resulted in the best accuracy of 83%. Similarly, Hiremath [14] et al. used NLP techniques on drug review data to determine a patient's sentiment towards a particular drug (positive, negative, or neutral). In this case as well, the SVM algorithm was found to perform better in terms of accuracy and precision. Hossain et. al. [12] proposed a drug recommendation framework that applies a sentimental measurement approach to drug reviews, generating ratings on drugs based on different features. Three machine learning algorithms, Decision Tree, K-Nearest Neighbors, and Linear Support Vector Classifier, are used to achieve the rating generation task, and their performances are compared to select the LinearSVC model as the optimal algorithm. The study shows that the sentimental attributes contribute greatly to the prediction of drug rating. Joshi et. al. [13] investigated the use of online reviews to predict medical conditions based on users' opinions of pharmaceutical drugs. Six different supervised machine learning classifiers were deployed to find the most efficient model, and Linear Support Vector Classifier (SVC) proved to perform best in terms of accuracy and training time.

Deep learning models can often achieve higher accuracy than traditional machine learning techniques, especially when dealing with complex data [15]. Min [16] proposed a weakly supervised mechanism (WSM) for adverse drug reactions identification in drug reviews on health forums. The proposed model, WSM-CNN-LSTM, combines the strengths of Convolutional Neural Network (CNN) and Bi-directional Long Short-term Memory (Bi-LSTM). The main advantage of this model is that it requires only a small quantity of labeled examples to attain optimal performance, which reduces the manual data-labelling requirements. Yadav et.al. [17] proposed a benchmark setup for analyzing sentiment in medical commu-

nities on social media platforms. The paper provides a fine-grained annotation scheme to capture sentiment in medical settings, using a deep convolutional neural network-based classification framework to predict possible medical sentiment categories.

III. Methodology

The work described in this paper was conducted by following the Knowledge Discovery in Databases (KDD) methodology. The proposed models were trained and tested using the "UCI ML Drug Review" dataset [1] which was designed to evaluate drug experience sentiment across various aspects, such as effectiveness or potential side effects. The dataset was obtained by crawling the pharmaceutical review website "Drugs.com" and gathering various attributes for each product such as the name of the targeted condition, a numerical 10 star rating, a textual patient review, and the number of patients who found the review useful. The dataset is split by 75% into a training set and 25% into a test set. It contains a total of 215,063 reviews for 6,345 individual drugs.

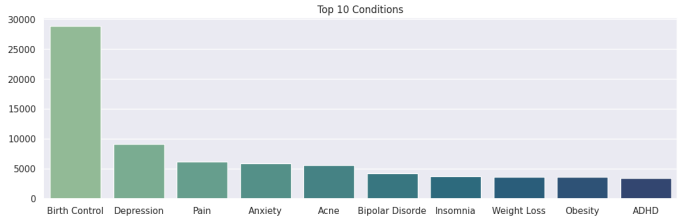

Fig. 1. Distribution of Top 10 Medical Conditions

The first step described by the KDD methodology is data collection. We performed an initial analysis of the data contained in the imported dataset using data visualization techniques. Figure 1 shows the ten topmost medical conditions targeted by the drugs included in the dataset. This information is relevant in the context of NLP because of the specific terminology associated with each affection. If the dataset is biased towards a single category, such as Birth Control in this case, the resulting model may be skewed towards that class and may not generalize well on new data. Another factor to consider is that conditions with shared terminologies, such as depression and anxiety in this case, can make it more difficult for the model to distinguish between the two categories and may lead to an overall decrease in accuraccy and precision.

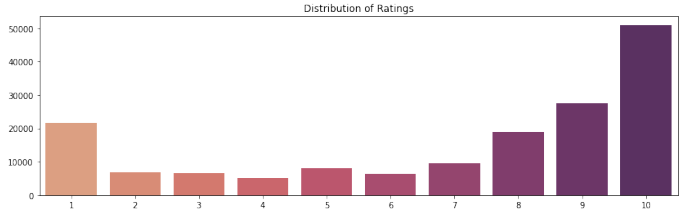

Fig. 2. Distribution of Drug Reviews by Rating Score

Figure 2 shows the distribution of drug reviews as categorized by the rating of each drug. The graph shows that the distribution is not balanced and there are several outliers with a significantly higher number of reviews, such as 1, 9 and 10. Unbalanced data can be a significant challenge in NLP because it can skew the performance of the machine learning algorithms and lead to overfitting. This is because the model is more likely to be biased towards the majority class, and hence, it may not learn to recognize the features that are unique to the minority classes.

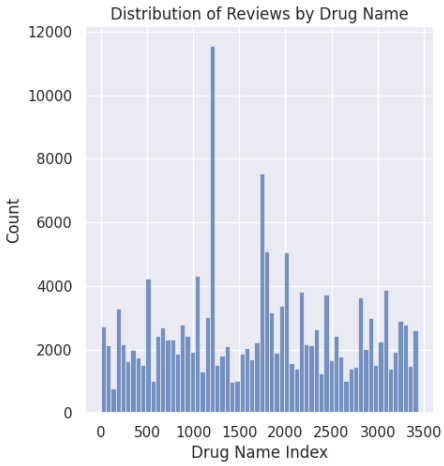

Fig. 3. Distribution of Drug Reviews by Drug Name

Figure 3 shows the distribution of drug reviews as categorized by the name of each drug. In this case we can see one outlier standing out with six times more data samples than the average. This can also introduce bias in the model's performance, as there is a risk of overfitting the majority type of reviews while performing poorly on other types of drugs.

For data cleaning, the reviews were normalized by converting all text to a standard format, applying lower case and removing punctuation and other non-alphabetic characters. Then, stop words were removed from the text to reduce noise and improve the accuracy of the prediction. We excluded the word 'not' from the list of stop words since negation is important in clinical texts, as it usually reverses the polarity of statements [2]. Once the data was cleaned, it was transformed into a format that can be used for analysis. We used the Bag of Words (BoW) model to represent the reviews, converting the text into an array which measures the frequency of each word. A limit of 500 was set on the number of features for the BoW model, in order to improve training performance and eliminate words with low frequencies which do not contribute to the overall accuracy of the prediction.

For the KNN model we used undersampling to balance the class distribution in the dataset by reducing the number of instances in the majority class. For the Naive Bayes and Random Forests models we addressed data imbalance by applying the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE works by creating synthetic samples of

the minority class through interpolation between the existing minority class examples.

The next step was to perform the data mining procedure. The processed dataset was used as input for three different machine learning models which were trained to predict the overall rating score of each review.

First, we applied Random Forests classification to categorize each review in one of the ten groups. This algorithm creates multiple decision trees through the random selection of data samples and subsets of features. Each tree is then trained on a subset of the data, and their individual predictions are combined to generate a final prediction. Random forests are suitable for NLP tasks since they are able to handle a large number of features, which is often the case in textual data. We used entropy as the criterion for building the decision trees, which has the advantage of producing more balanced trees while reducing overfitting.

The second algorithm tested was Naive Bayes, which is a probabilistic model using Bayes' theorem to determine the probability of a specific text belonging to a particular category or class. Naive Bayes is a simple and effective algorithm for NLP tasks, especially when there are numerous features and a small dataset. In our work we applied a variant of the algorithm called Gaussian Naive Bayes, which assumes that the distribution of words and features follows a Gaussian distribution for each class.

Lastly, we tested the clustering model K-Nearest Neighbours (KNN). The KNN algorithm falls under the category of supervised learning techniques, and it works by making predictions on the classification of a new data point based on how close it is to other data points in a feature space. In order to calculate the distance between points we used a metric called Minkowski distance, which is a generalization of the Manhattan and Euclidian distances.

IV. Evaluation

The first step in our evaluation was to reduce the problem of predicting drug ratings to a binary classification task. Each drug review was labeled as positive if its rating exceeded 5 points or negative for all scores below this value. Table I shows the results for each of the three models. The best performance was achieved by Random Forest, with an accuracy of over 90%.

TABLE I
Binary Classification Scores

Metric	*Random Forest*	*Naive Bayes*	*KNN*
Accuracy	0.924	0.705	0.884
Precision	0.842	0.737	0.832
Recall	0.945	0.629	0.789
F1 score	0.925	0.679	0.843

After evaluating the binary classification, we extended the problem to multiclass classification. We considered the Accuracy, Precision, Recall and F1 score as our parameters for the evaluation. Table II shows the results for each of the three models. From the given results, it can be clearly seen that Random Forest outperformed the other two models.

TABLE II
Multiclass Classification Scores

Metric	*Random Forest*	*Naive Bayes*	*KNN*
Accuracy	0.758	0.621	0.606
Precision	0.813	0.582	0.522
Recall	0.758	0.601	0.561
F1 score	0.751	0.564	0.504

As the data was highly imbalanced, we compared different sampling techniques to balance the class distribution and improve model performance: undersampling, oversampling and SMOTE (Synthetic Minority Oversampling Technique) . Table III shows the results for each of the three sampling methods. Oversampling techniques proved to be more efficient than undersampling, with the best results achieved by the classic oversampling method. Oversampling is generally used when the minority class is too small to produce relevant correlations, such as the present case when ratings from '2' to '6' have less than 10000 samples, compared to the majority class '10' which has around 50000, as shown in Figure 2. By comparison, undersampling works better when there is a majority class with a large number of instances but minority classes have enough samples to ensure a good prediction. The difference between SMOTE and oversampling is that oversampling balances class distribution by duplicating existing instances, while SMOTE performs interpolation between existing samples. SMOTE does not work well if the data is highly nonlinear and complex, such as the present case, which is why we see a lower performance compared to the classic oversampling method.

TABLE III
Sampling Methods For Random Forest

Metric	*Undersampling*	*Oversampling*	*SMOTE*
Accuracy	0.638	0.758	0.755
Precision	0.729	0.813	0.760
Recall	0.638	0.758	0.755
F1 score	0.625	0.751	0.756

Figure 4 shows the training and validation scores for the Random Forest classifier as evaluated on the training data. Generally, the validation score is expected to improve as the model learns to generalize better on new data but eventually starts to degrade once the point of overfitting is reached. The training score shows the performance of the model on the training data, which is why it is expected to be higher and more stable than the validation score. The graph shows that in this case the validation score improves considerably over time while the training score remains fairly constant at around 90%. This indicates that the model has good generalization capabilities and does not overfit the training data.

Overall we found that the Random Forest classifier performs best on both the binary and multiclass classification problems.

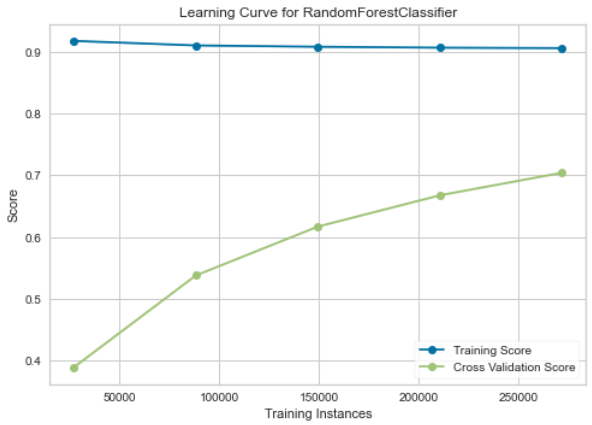

Fig. 4. Training and Validation Score for Random Forest Classifier

The most probable reason is that Random Forest models are able to handle a large number of features and capture complex nonlinear relationships between the features and the target variable. By contrast, Naive Bayes is considered to be more effective for small datasets and assumes an independent relationship between features. This can limit its accuracy, particularly when there are strong correlations between data attributes. KNN can capture local patterns in the data and does not make strong assumptions about the underlying distribution of the dataset. For this reason it may not perform well when there are many irrelevant or noisy features. Since our dataset contained a large number of features, associating textual data with other correlated attributes such as the corresponding medical condition and the name of the drug, the Random Forests model was the most suited for capturing the associations between features and predicting the final scores.

V. Conclusions and Future Work

Predicting drug efficacy based on patient reviews can be a valuable method for improving the research and development of new and existing drugs. Online patient reviews provide real-world data on drug efficacy in a variety of patient populations and settings. This information can complement data from clinical trials and help researchers understand how a drug performs in the real world. In this paper we applied data analysis to extract relevant information from patient reviews and through the use of data mining techniques we built models capable of predicting the efficacy of different drugs based on a 1-10 scale rating. We compared three different machine learning models - Random Forests, Naive Bayes and K-Nearest Neighbours - and their performance was evaluated on a real-world dataset: "UCI ML Drug Review". We tested the models on two separate problems: binary classification of the sentiment associated with each review (positive or negative) and multi-class classification of the drug rating score (on a 1-10 scale). Results showed that Random Forests performed best in both scenarios, achieving an accuracy of over 90% for binary classification and 75% for multi-class prediction.

Future research might focus on the integration of other forms of patient-generated information, including wearable devices and social media, thus further enhancing the accuracy of drug efficacy prediction models. Another possible research direction is the use deep learning models, which, as opposed to classic machine learning algorithms, have the ability to capture complex relationships such as the interactions between patient characteristics and drug effects. Deep learning models can also be updated and improved as more data becomes available, which can lead to better predictions and insights into drug efficacy over time.

References

[1] Gräßer, F., Kallumadi, S., Malberg, H., & 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).

[2] Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., & Rudzicz, F. (2019). A survey of word embeddings for clinical text. Journal of Biomedical Informatics, 100, 100057.

[3] Asghar, M. Z., Ahmad, S., Qasim, M., Zahra, S. R., & Kundi, F. M. (2016). SentiHealth: creating health-related sentiment lexicon using hybrid approach. SpringerPlus, 5, 1-23.

[4] Liu, S., & Lee, I. (2019). Extracting features with medical sentiment lexicon and position encoding for drug reviews. Health information science and systems, 7, 1-10.

[5] Saad, E., Din, S., Jamil, R., Rustam, F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Determining the efficiency of drugs under special conditions from users' reviews on healthcare web forums. IEEE Access, 9, 85721-85737.

[6] Gräßer, F., Kallumadi, S., Malberg, H., & 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).

[7] Han, Y., Liu, M., & Jing, W. (2020). Aspect-level drug reviews sentiment analysis based on double BiGRU and knowledge transfer. IEEE Access, 8, 21314-21325.

[8] Imani, M., & Noferesti, S. (2022). Aspect extraction and classification for sentiment analysis in drug reviews. Journal of Intelligent Information Systems, 1-21.

[9] Sweidan, A. H., El-Bendary, N., & Al-Feel, H. (2021). Sentence-level aspect-based sentiment analysis for classifying adverse drug reactions (ADRs) using hybrid ontology-XLNet transfer learning. IEEE Access, 9, 90828-90846.

[10] Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. International Journal, 2(6), 282-292.

[11] Marthin, P., & Duygu, İ. Ç. E. N. (2020). Application of natural language processing with supervised machine learning techniques to predict the overall drugs performance. AJIT-e: Academic Journal of Information Technology, 11(40), 8-23.

[12] Hossain, M. D., Azam, M. S., Ali, M. J., & Sabit, H. (2020, December). Drugs rating generation and recommendation from sentiment analysis of drug reviews using machine learning. In 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE) (pp. 1-6). IEEE.

[13] Joshi, S., & Abdelfattah, E. (2021, May). Multi-class text classification using machine learning models for online drug reviews. In 2021 IEEE World AI IoT Congress (AIIoT) (pp. 0262-0267). IEEE.

[14] Hiremath, B. N., & Patil, M. M. (2022). Enhancing optimized personalized therapy in clinical decision support system using natural language processing. Journal of King Saud University-Computer and Information Sciences, 34(6), 2840-2848.

[15] Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. Journal of Biomedical Informatics, 110, 103539.

[16] Min, Z. (2019, March). Drugs reviews sentiment analysis using weakly supervised model. In 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 332-336). IEEE.

[17] Yadav, S., Ekbal, A., Saha, S., & 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).
[18] Mishra, A., Malviya, A., & Aggarwal, S. (2015, November). Towards automatic pharmacovigilance: analysing patient reviews and sentiment on oncological drugs. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW) (pp. 1402-1409). IEEE.
[19] Goeuriot, L., Na, J. C., Min Kyaing, W. Y., Khoo, C., Chang, Y. K., Theng, Y. L., & Kim, J. J. (2012, January). Sentiment lexicons for health-related opinion mining. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (pp. 219-226)

VI. Assignment

Group Number: 34
Presentation: https://youtu.be/D8w1bcVHgxw
Project: https://github.com/oanad1/drug-efficacy-prediction