Published in final edited form as:
IEEE Int Conf Healthc Inform. 2022 June ; 2022: 163–169. doi:10.1109/ichi54592.2022.00035.

Classifying Drug Ratings Using User Reviews with Transformer-Based Language Models

Akhil Shiju,
Department of Biological Sciences, Florida State University, Tallahassee, Florida, USA

Zhe He
School of Information, Florida State University, Tallahassee, Florida, USA

Abstract

Drug review websites such as Drugs.com provide users' textual reviews and numeric ratings of drugs. These reviews along with the ratings are used for the consumers for choosing a drug. However, the numeric ratings may not always be consistent with text reviews and purely relying on the rating score for finding positive/negative reviews may not be reliable. Automatic classification of user ratings based on textual review can create a more reliable rating for drugs. In this project, we built classification models to classify drug review ratings using textual reviews with traditional machine learning and deep learning models. Traditional machine learning models including Random Forest and Naive Bayesian classifiers were built using TF-IDF features as input. Also, transformer-based neural network models including BERT, Bio_ClinicalBERT, RoBERTa, XLNet, ELECTRA, and ALBERT were built using the raw text as input. Overall, Bio_ClinicalBERT model outperformed the other models with an overall accuracy of 87%. We further identified concepts of the Unified Medical Language System (UMLS) from the postings and analyzed their semantic types stratified by class types. This research demonstrated that transformer-based models can be used to classify drug reviews based solely on textual reviews.

Keywords

Drug; Classification; Natural Language Processing

I. Introduction

The evaluation of the efficacy and safety of drugs heavily relies on randomized controlled trials with rigorous inclusion and exclusion criteria [1]. However, such processes are limited to a small number of individuals enrolled in the study and are constrained to participants in the target population who meet possibly restrictive eligibility criteria, limiting the population representativeness and subsequent study generalizability [2], [3]. The ramifications of these acclimations could potentially have resulted in the overestimation of the efficacy of the product and misidentification of adverse events/side effects in the diverse population [4]. To counter such issues, approaches such as post-marketing drug surveillance have been

aks19m@my.fsu.edu .

introduced to optimize the safety of the drug after its regulatory approval and mass production [5]. There are two major forms of post-marketing drug surveillance. Some are formed by government regulators such as the Vaccine Adverse Event Reporting System (VAERS) by the United States Food and Drug Administration [6] or the Yellow Card Scheme by the United Kingdom Medicines and Healthcare Products Regulatory Agency [7]. Also, public/private organizations have a system to monitor drug side-effects such as RADAR (the Research on Adverse Drug Events and Reports) [8]. Existing methods for identifying adverse events typically focused on analyzing molecular drug composition, [9] query logs, [10] VAERS records, [11] or clinical notes in the medical records [12] but did not analyze specifically the sentiment of the consumers using their reviews of the drug [13]. The application of post-market drug surveillance has been successfully applied in the identification of adverse events through safety reports by the introduction of deep learning-based methods including the extraction of temporal events, the procedure performed, and social circumstance [14]. In the era of Web 2.0, the Internet has opened up new pathways to obtain information directly from consumers about their drug reviews in an elaborative format. Publicly available information on the Internet offers an easily attainable resource that could be leveraged to gain a deep understanding of the drug reviews by the users. Entire user reviews are fully available on drug review websites, on which users can comment on their personal experiences of the drugs they have taken for a specific condition. Unlike many other forms of medical data, this information is not filtered through medical professionals. Since these reviews are given by anonymous users, there is no risk of patient health record violation for confidentiality. These reviews contain a plethora of information regarding individual experiences associated with the drugs such as symptoms, adverse events, and interactions with other drugs. Such reviews have also contained an extensive amount of user sentiment related to a particular condition, which could be leveraged to detect the side effects and efficacy of drugs [15].

However, many barriers exist in the extraction of sentiment from these online medical reviews. For instance, user reviews of drugs in such online forms are typically unconventional and most reviewers lack medical knowledge, posing barriers for extracting meaningful information from them. In addition, many review websites use some form of numerical rating that serves the role of quantifying such a sentiment, but they do not provide a clear guideline for giving a certain numeric rating. As such, these review websites may have introduced biases as individual users may have different perception as to what a high score entails versus what would constitute a low score. Users tend to reduce the effort required in reporting values by rating all qualities as highly important, thus resulting in overly positive ratings [16]. This could lead to an unintended positive view of the overrated drugs by the general public, albeit less effective for certain population subgroups. Prior research has found that web-based reviews have the potential to be viewed as an applicable source of information for analysis, but review scoring biases may exist. For example, addictive drugs have been observed to be typically numerically highly rated in comparison to other drugs which have treated the same condition, even if these additive drugs underperformed based on experience [17]. Prediction of user review score directly from user input may provide a method to limit this issue [18].

The application of machine learning, especially through transformer-based language models pre-trained with an enormous amount of data, offers a unique approach to classify textual information [19]. In this project, we evaluated the feasibility of leveraging machine learning and natural language processing to classify user ratings based on their textual review. To provide some interpretability of the classification results, we used an interpretation tool called Eli5 to highlight phrases in the text that have a positive or negative impact on the classification results. We further used QuickUMLS to identify semantic types and analyzed their associations with the classifications.

II. Methods

A. Dataset Preparation

We obtained the dataset from the UCI Machine Learning Repository [20]. These instances were collected from Drugs.com using Beautiful Soup. The dataset used for this study consists of user drug reviews, drug names, related medical conditions, and a 10-point rating. The rating were integer values ranging from 1 to 10 with 10 being the highest possible rating. Table 1 shows example records of the dataset. Fig. 1 shows the distribution of reviews by ratings. The ratings were shown to be skewed to the left to suggest that most drugs received a relatively high score. Prior analysis of this dataset focused primarily on the sentiment analysis [21] and classification of reviews using an n-gram technique with unequal classes, thus skewing accuracy [22]. Neither was there an emphasis on the error analysis of the models. In total, the dataset consists of 215,063 instances. The numeric ratings had a mean of 7.00 with a standard deviation of 3.27. There are 836 classified medical conditions in the dataset.

B. Review Rating Classification

Since the primary focus of this study was to classify textual reviews, the data was broken down into two rating groups using a median of the ratings: ratings 8 or above were considered *above average*, and below 8 were considered as *below average*. The binary classification was chosen over multiple classes since the objective of this project was to minimize user subjectivity for numeric rating, and based on the distribution of the review rating scores, the introduction of additional classes may result in less distinguishable features within the groups. Instances in which the reviews contained more than 514 tokens were removed from the study due to the input size limit of the transformer-based language models.

The common methodology for transfer learning has been applied through the application of pre-training on a large unannotated corpus that was capable of understanding the composition of the data type such as patterns in the language. This process could be considered as self-supervised learning. This pre-trained model is then followed by the fine-tuning process which focused on the training on an application-specific dataset.

BERT: Some common language models are pre-trained by predicting the next word in a sequence, but Bidirectional encoder representation from transformer (BERT) looked at bidirectional predicting context masked intermediate text tokens in the pretraining from

Wikipedia and BookCorpus and next sentence prediction. Bert-base-uncased was used for this project [23].

Bio_ClinicalBERT: The BERT model has been pre-trained with a medical corpus from publicly available data from PubMed, PMC, and MIMIC III clinical notes [24].

ALBERT: A Lite BERT (ALBERT) is a model which focused on being a less memory-heavy and faster version of BERT through the separation of the word embedding into two matrixes and by cross-layer parameter sharing [25]. Albert-base-v2 was used for this model.

RoBERTa: Robustly Optimized BERT Approach (RoBERTa) has been considered a pretraining model that eliminates the next sentence prediction task and adapts a novel approach of dynamic masking which randomized the masked token between training epochs [26]. RoBERTa outperformed BERT on multiple benchmarks such as GLUE, RACE, and SQuAD. Roberta-base was the model selected for this project.

XLNet: As a more computationally expensive model, the Generalized Auto-Regressive model (XLNet) implemented a system where it applies an autoencoder language model [27].

ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) replaced the masked language task with a generator and pre-trains the model to identify which token has been replaced [28]. The Electra-base-discriminator was used for this project.

We split the dataset into a training set (60%), a validation set (20%), and the test set (20%). These datasets were further classified into lists which were then converted into Transformer datasets that could be trained by a neural network to generate a model.

We constructed these transformer-based text classification models utilizing the Huggingface transformers using the Python k-train pipeline wrapper class for text classification. The models used for this project consisted of Bio_ClinicalBERT, ELECTRA, RoBERTa, XLNet, ALBERT, and BERT. The parameter included a 514 max token length, a 5e^-5 learning rate, and a batch size of 6. The train test dataset was fed into the neural network trained to minimize validation data loss. After the training was completed, a confusion matrix of the test data was generated to determine F1 scores for the classes (threshold of 0.5) and the accuracy in comparison to the user ratings.

As a baseline approach for evaluating the transformer-based models, bag-of-words (BOW) models were constructed based on term frequency and inverse document frequency (TF-IDF). The textual reviews were converted into a bag of words representation. Afterward, a term TF-IDF score matrix was computed for the bag of words representation. We trained and evaluated a Random Forest classifier and a Naïve-Bayes classifier with the BOW features.

The test data was stratified for the top 10 conditions based on the test data user reviews. The transformer models were then used to classify each of the different conditions to determine condition-specific F1 score and accuracy. The overall workflow is outlined in Fig. 2.

C. Model Interpretation

After the best-performing transformer model was selected, to provide some interpretability for the model, the LIME algorithm [29] implemented as the TextExplainer in Eli5 package was applied to the model to highlight important terms for classification. Eli5 has been used to understand why a certain classification is made by the model by identifying important features such as highlighting significant text features [30]. This is accomplished by inspecting the model parameters to discover the global implications. This was performed for all reviews to establish some sense of how the model performed these classifications. Top scores were also computed through the Eli5 metrics.

D. Error Analysis

We first examined the words highlighted by Eli5 in false positive and false negative instances. Then, an analysis of the potential relationship between false positives, false negatives, true positives, true negatives from the best overall performing models was conducted by analyzing the occurrences of certain semantic types of the Unified Medical Language System (UMLS) Metathesaurus, which links terms to biomedical concepts [31]. We would like to see whether certain error types had deviation in the semantic types present in the review in comparison to the other conditional cases. This was conducted using the QuickUMLS package, an unsupervised tool for biomedical term extraction using simstring [32]. We chose top 10 semantic types that were most prevalent in the dataset and had some medical significance, including Sign or Symptom, Disease or Syndrome, Organism Function, Pathologic Function, Body Substance, Body Location, or Region, Body Part, Organ, or Organ Component, and Health Care Activity, Antibiotics, and Physiologic Function. Only instances with a 1.0 Jaccard similarity were retained, and the best matching CUIs were selected. After the semantic types were extracted for all the reviews, the means were calculated by true positive, true negative, false positive and false negative (class type). A one-way ANOVA was employed to determine whether there was a significant difference based on the mean value of the number of concepts of a certain semantic type per post across different class types.

III. Results

Overall, the model generated by the Bio_ClinicalBERT and ELECTRA outperformed the other models on a variety of metrics as displayed in Table 2. The BOW models showed lower accuracy compared to the others. XLNet had the longest training time compared to the other models. Table 3 provides the condition-specific statistics for the top 10 conditions. The ratings of the reviews pertaining to the Birth Control, Depression, Pain drugs were classified with high accuracy than the ratings of the drugs for other conditions. The conditions with lower instances had lower accuracy than the conditions with higher instances. However, there are many notable deviations present such as the pain and obesity models' lower accuracy or the higher accuracy for the ADHD model.

A. Error Analysis

The results generated in Table 4 are produced by applying the Eli5 toolkit to the Bio_ClinicalBERT trained model. There is a clear relationship between the words highlighted and the rating group classified by the model. Terms highlighted in green supports the classification generated by the model, while terms generated in red oppose the predictions. The shade of the color represents the level of importance at which a word contributes to the classification. Phrases related to side effects were typically highlighted as *below-average* features such as being a "bit moody" or "sore". The positive effects of the drug were highlighted as *above-average* features such as "my pain almost totally disappeared". Specific highlighted terms by the ELi5 metric could potentially be subjected to an incorrect sentiment association. For example, the phrase "my cramps disappeared" in Table 4 for the false negative adverse event review was shown to support the prediction of a negative below average feature. However, this phrase would usually have a positive connotation associated with it, unlike what is suggested by the model. In general, the major reason for a wrong prediction by the model was primarily due to the presence of a mixture of positive and negative words present in the text. This could have resulted from the presence of multiple medications, changes in the effectiveness of the medication over time, the extent of the medicinal effects, or treatment experience which could work in both sentimental directions to result in false positives or false negatives. Adverse events, in general, could result in false positive or false negatives depending on the extent to which these adverse events concerned someone. In addition, if a classification error occurred, a lower number of adverse events tended to be classified as false positive versus false negative. Overall, the interpretability of misclassifications, through the Eli5 tool kit revealed an important aspect of how the model used specific keywords.

Regarding the correlation analysis of semantics types with class types, most semantic types were found to differ significantly by class types with a p-value < 0.05 based on the results shown in Table 5. This indicate that these semantic types did play a role in helping the model classify reviews. Physiologic Function for the Bio_ClinicalBERT model and Clinical Drug semantic type for both models were found to be insignificant. This suggests that both models tend not to heavily rely on the name of the clinical drug in predicting a score but could also be due to the lack of clinical drug names present in the user reviews. This idea is further supported based on the results of the ELi5 which shows many clinical drugs highlighted less impactful (lighter) to the classification than other terms in general. Based on results of the ELECTRA model, the average number of UMLS concepts of most semantic types (e.g., Sign or Symptom) in true negative instances is greater than that of true positive instances. The average numbers of concepts of most semantic types in the false positive and false negative instances are between that of true positive and true negative instances. This could explain why the model made false classifications.

However, the Bio_ClinicalBERT model has some more significant deviation from the most common class distribution in the ELECTRA model. Significant deviation from this class type distribution in the Bio_ClinicalBERT model occur for Disease or Syndrome, Organism Function, Pathologic Function, Body region or Location. Both models share similarities in their ranking of Sign or Symptom, Body Substance, Body Part, Organ, or Organ Component, Health Care Activity.

IV. Discussion

Transformer models provide an automated, fast, and economic system to classify the sentiment of reviews from individuals for specific medications. Furthermore, transformer models' capability to generate a suggestion of a score solely based on user reviews can be utilized as a point of comparison to user-generated reviews. Accurately rating drug reviews can help consumers identify positive or negative reviews without having to sort through reviews which lack standardization in scoring. In a clinical study, this could potentially contribute towards advancing a conversation with the reviewers to further investigate the cause for such variations.

In addition, Eli5 is an easy tool to understand which noteworthy terms contribute to the model, and potentially reviewers relied on providing more clarity on the logic behind the score. The identification of significant term contributors through the Eli5 metrics could hint at factors such as adverse events that are important in post-market drug surveillance. The binary classification approach of Bio_ClinicalBERT and other transformer models could aid in potentially finding negative drug reviews in data that lacks a numeric score. This filtration of reviews delivers a vital step to simplify the process in the identification of adverse events, side effects, and possible medical interactions. A fast-paced system sentiment score prediction attests to its impact in analyzing large social media drug datasets providing a manageable tool to separate reviews into separate classes. The classified social media data can then be adopted for different purposes such as topic modeling by sentiment types.

In this study, we built multiple classification models to classify drug review rating using the review text. Afterward, the Eli5 toolkit was applied to explain the models' classification by highlighting the words that positively or negatively impacted the classification result. Informed by these experiments, it was clear that the consumers' online drug reviews contain a vast quantity of information related to the sentiment expressed by the user. Transformer-based models have the potential to serve as a methodology to discriminate text reviews. Overall, this research outlined a potential process of standardizing drug review rating, limiting user input and reducing the subjective effects which are often distorted in rating systems [33].

When comparing different transformer-based models, the Bio_ClinicalBERT and ELECTRA models outperformed the other models when the same amount of information was present. One possible reason for Bio_ClinicalBERT to outperform other models is that Bio_ClinicalBERT model was pre-trained with biomedical texts which were topically related to the drug reviews [34]. According to the error analysis, ELECTRA model followed the pattern that the average numbers of concepts of most semantic types per post in true negative posts were greater than those in true positives; and those of false positive and false negative instances falling between those of true cases. This is further cemented by the fact that the ELi5 toolkit highlighted these terms as more important contributors in general. The Bio_ClinicalBERT model tends to evade this classification for certain semantic types as previously stated in the error analysis. Bio_ClinicalBERT trained with biomedical text allowed it to find more intricate relationships among the terms, allowing it to reach a better prediction accuracy. However, Bio_ClinicalBERT and ELECTRA did significantly follow a similar pattern for the semantic types such as Sign or Symptom, Body substance, Body part organ or Organ component, and Health care activity. As it is clear that both models relied significantly on these factors to decern the sentiment of a user review, it is likely that user also weighted these factors higher than other semantic types when deciding their rating of the drugs. As such, higher number of possible adverse events (sign or symptoms), the need for more possible medical interventions (healthcare activities), and more reference to bodily fluids and organs (body substances and body part organ or organ component) tend to result in a lower rating.

Type 1 and type 2 errors in the model classification occurred due to a plethora of reasons which could be broadly classified into model-based error or user-based errors. Model based error previously mentioned were due to the presence of multiple different positive or negative aspects in the text confounding the prediction generated. User-based errors are typically knowledge gaps for the score criteria or a certain level of subjectivity of the scoring such as considering the score of 1 as the best possible rating when according to the guidelines it is the worst.

Although there have been other studies which have built deep learning models for sentiment analysis of drug reviews, this study's novelty arose from achieving high metrics with a binary classification model and the application of eli5 and QuickUMLS. In comparison to other studies, BIO_CLINICALBERT achieved similar results, under different class sizes [35].

A. Limitations and Future Work

Although this model was able to successfully classify reviews in a binary system, the ability for large class identification is still unknown and warrants further investigation. One important issue with many transformer models was the issue of over-fitting. In addition, many transformer models such as XLNet are computationally expensive which may require in a long training time. Additional research will concentrate on the utilization of transformer models on non-scored-based social media data. Also, another area of focus could be to expand this model for multi-class identification as this may be more advantageous in the determination of highly negative reviews and understanding probable reasons for model inaccuracy.

V. Conclusions

This study presents the construction of transformer-based models for the classification of drug reviews from drugs.com. The most successful model in this project was the Bio_ClinicalBERT model with the highest F1 score. Overall, the transformer models outperformed the traditional machine learning models using bag-of-words features. These binary transformer models tended to be effective at discerning optimistic reviews from negative reviews.

Acknowledgment

This study was partially supported by the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Number R21AG061431; and in part by Florida State University-University of Florida Clinical and Translational Science Award funded by National Center for Advancing Translational Sciences under Award Number UL1TR001427.

References

[1]. Califf RM, “Characteristics of Clinical Trials Registered in ClinicalTrials.gov, 2007–2010,” JAMA, vol. 307, no. 17, p. 1838, May 2012, doi: 10.1001/jama.2012.3424. [PubMed: 22550198]

[2]. Farmer KC, “Methods for measuring and monitoring medication regimen adherence in clinical trials and clinical practice,” Clin. Ther, vol. 21, no. 6, pp. 1074–1090, Jun. 1999, doi: 10.1016/S0149-2918(99)80026-5. [PubMed: 10440628]

[3]. He Z et al. , “Clinical Trial Generalizability Assessment in the Big Data Era: A Review,” Clin. Transl. Sci, vol. 13, no. 4, pp. 675–684, Jul. 2020, doi: 10.1111/cts.12764. [PubMed: 32058639]

[4]. Mills EJ et al. , “Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors,” Lancet Oncol, vol. 7, no. 2, pp. 141–148, Feb. 2006, doi: 10.1016/S1470-2045(06)70576-9. [PubMed: 16455478]

[5]. Crombie I, “The role of record linkage in post-marketing drug surveillance.,” Br. J. Clin. Pharmacol, vol. 22, no. S1, pp. 77S–82S, Feb. 1986, doi: 10.1111/j.1365-2125.1986.tb02987.x. [PubMed: 3567036]

[6]. Shimabukuro TT, Nguyen M, Martin D, and DeStefano F, “Safety monitoring in the Vaccine Adverse Event Reporting System (VAERS),” Vaccine, vol. 33, no. 36, pp. 4398–4405, Aug. 2015, doi: 10.1016/j.vaccine.2015.07.035. [PubMed: 26209838]

[7]. O’Donovan B, Rodgers RM, Cox AR, and Krska J, “Making medicines safer: analysis of patient reports to the UK’s Yellow Card Scheme,” Expert Opin. Drug Saf, vol. 18, no. 12, pp. 1237–1243, Dec. 2019, doi: 10.1080/14740338.2019.1669559. [PubMed: 31538503]

[8]. Yom-Tov E and Gabrilovich E, “Postmarket Drug Surveillance Without Trial Costs: Discovery of Adverse Drug Reactions Through Large-Scale Analysis of Web Search Queries,” J. Med. Internet Res, vol. 15, no. 6, p. e124, Jun. 2013, doi: 10.2196/jmir.2614. [PubMed: 23778053]

[9]. Dey S, Luo H, Fokoue A, Hu J, and Zhang P, “Predicting adverse drug reactions through interpretable deep learning framework,” BMC Bioinformatics, vol. 19, no. S21, p. 476, Dec. 2018, doi: 10.1186/s12859-018-2544-0. [PubMed: 30591036]

[10]. Ahmad F, Abbasi A, Kitchens B, Adjeroh DA, and Zeng D, “Deep Learning for Adverse Event Detection from Web Search,” IEEE Trans. Knowl. Data Eng, pp. 1–1, 2020, doi: 10.1109/TKDE.2020.3017786.

[11]. Moro PL, Arana J, Cano M, Lewis P, and Shimabukuro TT, “Deaths Reported to the Vaccine Adverse Event Reporting System, United States, 1997–2013,” Clin. Infect. Dis, vol. 61, no. 6, pp. 980–987, Sep. 2015, doi: 10.1093/cid/civ423. [PubMed: 26021988]

[12]. Dandala B, Joopudi V, and Devarakonda M, “Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks,” Drug Saf, vol. 42, no. 1, pp. 135–146, Jan. 2019, doi: 10.1007/s40264-018-0764-x. [PubMed: 30649738]

[13]. Kulldorff M, Davis RL, Kolczak M, Lewis E, Lieu T, and Platt R, “A Maximized Sequential Probability Ratio Test for Drug and Vaccine Safety Surveillance,” Seq. Anal, vol. 30, no. 1, pp. 58–78, Jan. 2011, doi: 10.1080/07474946.2011.539924.

[14]. Du J et al. , “Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning,” J. Am. Med. Inform. Assoc, vol. 28, no. 7, pp. 1393–1400, Jul. 2021, doi: 10.1093/jamia/ocab014. [PubMed: 33647938]

[15]. Dinh Thu, Chakraborty Goutam, and McGaugh Miriam, “Exploring Online Drug Reviews using Text Analytics, Sentiment Analysis and Data Mining Models”

[16]. Hino A and Imai R, “Ranking and Rating: Neglected Biases in Factor Analysis of Postmaterialist Values,” Int. J. Public Opin. Res, vol. 31, no. 2, pp. 368–381, Jun. 2019, doi: 10.1093/ijpor/edy007.

[17]. Tanabe P and Buschmann M, “A prospective study of ED pain management practices and the patient’s perspective,” J. Emerg. Nurs, vol. 25, no. 3, pp. 171–177, Jun. 1999, doi: 10.1016/S0099-1767(99)70200-X. [PubMed: 10346837]

[18]. Adusumalli S, Lee H, Hoi Q, Koo S-L, Tan IB, and Ng PC, “Assessment of Web-Based Consumer Reviews as a Resource for Drug Performance,” J. Med. Internet Res, vol. 17, no. 8, p. e211, Aug. 2015, doi: 10.2196/jmir.4396. [PubMed: 26319108]

[19]. Lewis DD, “Challenges in machine learning for text classification,” in Proceedings of the ninth annual conference on Computational learning theory - COLT ‘96, Desenzano del Garda, Italy, 1996, p. 1-ff. doi: 10.1145/238061.238062.

[20]. UCI. drug review dataset (drugs.com) data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

[21]. Vijayaraghavan S and Basu D, “Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms,” ArXiv200311643 Cs Stat, Mar. 2020, Accessed: Jan. 11, 2022. [Online]. Available: http://arxiv.org/abs/2003.11643

[22]. Gräßer F, Kallumadi S, Malberg H, and Zaunseder S, “Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning,” in Proceedings of the 2018 International Conference on Digital Health, Lyon France, Apr. 2018, pp. 121–125. doi: 10.1145/3194658.3194677.

[23]. Devlin J, Chang M-W, Lee K, and Toutanova K, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” ArXiv181004805 Cs, May 2019, Accessed: Jan. 11, 2022. [Online]. Available: http://arxiv.org/abs/1810.04805

[24]. Alsentzer E et al., “Publicly Available Clinical BERT Embeddings,” in Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, Minnesota, USA, 2019, pp. 72–78. doi: 10.18653/v1/W19-1909.

[25]. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, and Soricut R, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,” ArXiv190911942 Cs, Feb. 2020, Accessed: Jan. 11, 2022. [Online]. Available: http://arxiv.org/abs/1909.11942

[26]. Liu Y et al. , “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” ArXiv190711692 Cs, Jul. 2019, Accessed: Jan. 11, 2022. [Online]. Available: http://arxiv.org/abs/1907.11692

[27]. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, and Le QV, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” ArXiv190608237 Cs, Jan. 2020, Accessed: Jan. 11, 2022. [Online]. Available: http://arxiv.org/abs/1906.08237

[28]. Clark K, Luong M-T, Le QV, and Manning CD, “ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators,” ArXiv200310555 Cs, Mar. 2020, Accessed: Jan. 11, 2022. [Online]. Available: http://arxiv.org/abs/2003.10555

[29]. Ribeiro MT, Singh S, and Guestrin C, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[30]. Agarwal N and Das S, “Interpretable Machine Learning Tools: A Survey,” in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, Dec. 2020, pp. 1528–1534. doi: 10.1109/SSCI47803.2020.9308260.

[31]. Bodenreider O, “The Unified Medical Language System (UMLS): integrating biomedical terminology,” Nucleic Acids Res, vol. 32, no. 90001, pp. 267D–270, Jan. 2004, doi: 10.1093/nar/gkh061.

[32]. Soldaini Luca and Goharian Nazli, “QuickUMLS: a fast, unsupervised approach for medical concept extraction,” 2016, [Online]. Available: http://medir2016.imag.fr/data/MEDIR_2016_paper_16.pdf

[33]. Abou Taam M et al. , “Analysis of patients’ narratives posted on social media websites on benfluorex’s (Mediator ®) withdrawal in France,” J. Clin. Pharm. Ther, vol. 39, no. 1, pp. 53–55, Feb. 2014, doi: 10.1111/jcpt.12103. [PubMed: 24304185]

[34]. Pipalia K, Bhadja R, and Shukla M, “Comparative Analysis of Different Transformer Based Architectures Used in Sentiment Analysis,” in 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, Dec. 2020, pp. 411–415. doi: 10.1109/SMART50582.2020.9337081.

[35]. Colón-Ruiz C and Segura-Bedmar I, “Comparing deep learning architectures for sentiment analysis on drug reviews,” J. Biomed. Inform, vol. 110, p. 103539, Oct. 2020, doi: 10.1016/j.jbi.2020.103539. [PubMed: 32818665]

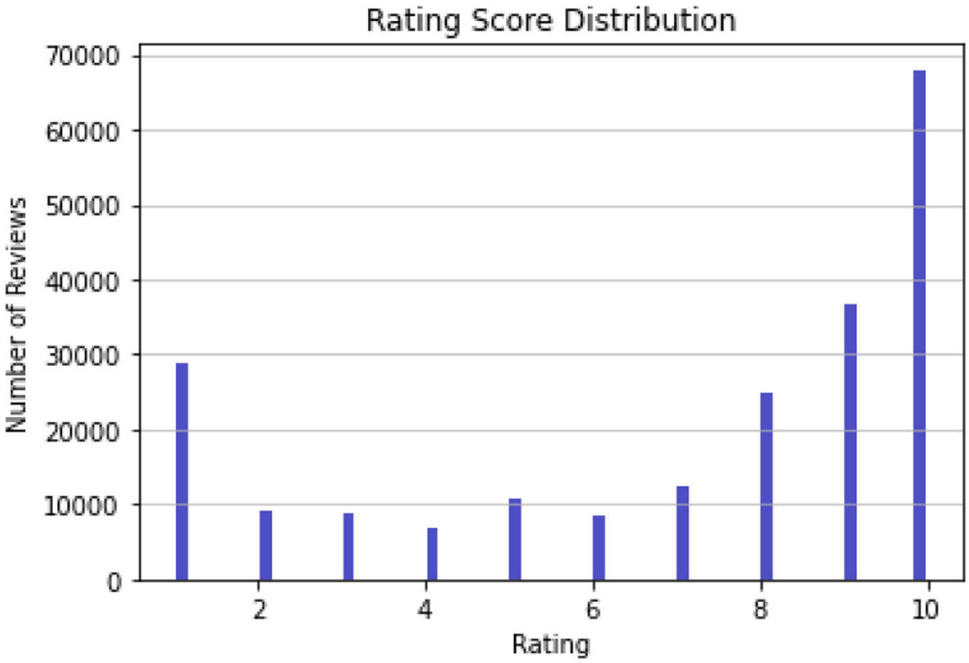

Fig 1.

Total number of reviews in the dataset.

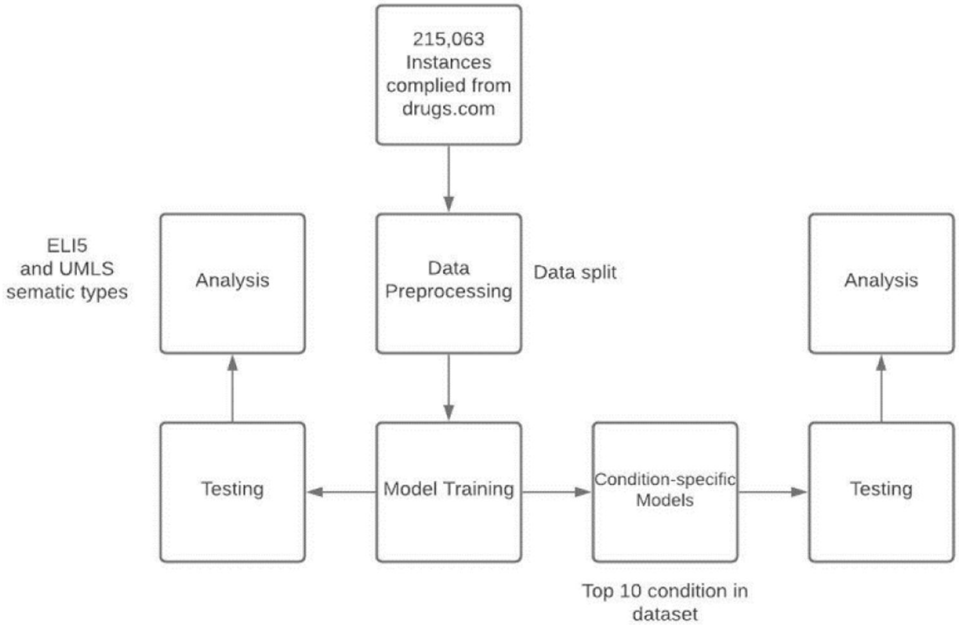

Fig 2.

The workflow of the project.

Table 1.

Two examples of a high-rating review versus a low-rating review with condition, drug name, and rating.

Drug Name	Condition	Review	Rating
Chantix	Smoking Cessation	I smoked for 50+ years. Took it for one week and that was it. I didn't think it was possible for me to quit. It has been 6 years now. Great product.	10.0
Excedrin	Migraine	Does not work for people sensitive to caffeine. I was jittery and nervous and queasy after using a single dose.	2.0

Table 2.

Overall condition validation from the test dataset for the minimized loss for the top-performing models.

Model	**Above Average F1**	**Below Average F1**	**Accuracy**
BERT	0.84	0.84	0.84
RoBERTa	0.83	0.83	0.84
XLNet	0.84	0.84	0.84
Bio_ClinicalBERT	**0.87**	**0.87**	**0.87**
ELECTRA	0.85	0.87	0.86
ALBERT	0.75	0.81	0.78
Random Forest (BOW)	0.77	0.45	0.68
Naïve Bayes (BOW)	0.76	0.03	0.61

Table 3.

Results of condition-specific classifications for the top 10 conditions.

Condition	Model	Above Average F1	Below Average F1	Accuracy
Birth Control	ELECTRA	0.89	0.94	0.92
	Bio_ClinicalBERT	0.86	0.92	0.90
Depression	ELECTRA	0.88	0.88	0.88
	Bio_ClinicalBERT	0.87	0.87	0.87
Pain	ELECTRA	0.83	0.81	0.82
	Bio_ClinicalBERT	0.85	0.83	0.84
Anxiety	ELECTRA	0.87	0.82	0.85
	Bio_ClinicalBERT	0.87	0.81	0.85
Acne	ELECTRA	0.90	0.88	0.89
	Bio_ClinicalBERT	0.86	0.84	0.85
Bipolar Disorder	ELECTRA	0.88	0.84	0.86
	Bio_ClinicalBERT	0.84	0.87	0.86
Insomnia	ELECTRA	0.82	0.86	0.84
	Bio_ClinicalBERT	0.82	0.84	0.83
Weight Loss	ELECTRA	0.87	0.80	0.85
	Bio_ClinicalBERT	0.89	0.79	0.86
Obesity	ELECTRA	0.82	0.77	0.80
	Bio_ClinicalBERT	0.85	0.79	0.83
ADHD	ELECTRA	0.86	0.88	0.87
	Bio_ClinicalBERT	0.88	0.87	0.87

Table 4.

Examples of false positives and false negatives in Bio_ClinicalBERT model with the important words highlighted in green (positively impacting the classification results) and red (negatively impacting the classification results) by the Eli5 toolkit.

Class type/reason	Review
False negative/multiple medications	for me, vyvanse has the "smoothest" feeling of the adhd medicines i have tried. i have found that concerta (methylphenidate) and focalin (dexmethylphenidate) create an anxious feeling. vyvanse does not make me feel this way. downside: it can be outrageously expensive.
False positive/Temporal effectiveness	When i first started lyrica, my pain almost totally disappeared. after about 3 weeks, my pain started returning. my tongue started to tingle and was sore.
False negative/adverse events	love this. cleared my skin up, made my period so light and my cramps disappear. i was a bit moody for the first month, but that went away.
False positive/ medication ineffectiveness	"i have cysteine stones...huge! passed 9 small stones within 30 mins after taking. and with very little pain in the uretha but doesn't help much with the ureter pain.
False positive/treatment experience	it gets the job done. tastes gross and i personally had a hard time keeping it down but i managed. it took about 2 hours for the first dose to kick in and I've been going since. took the 2nd dose an hour ago and almost clear!

Table 5.

The average mean number of semantic types reported for each class type based on the classification results of the overall Bio_ClinicalBERT and ELECTRA models. P-values were generated using a 1-way ANOVA test for each condition using semantic type as the independent variable.

Model	Semantic Types	Sign or Symptom	Disease or Syndrome	Antibiotics	Organism Function	Pathologic Function	Body Substance	Body Location or Region	Physiologic Function	Body Part, Organ, or Organ Component	Health Care Activity
ELECTRA	True Positive	1.167	0.595	0.013	0.522	0.577	0.079	0.256	0.12	0.579	0.382
	True Negative	2.317	0.925	0.023	0.867	0.879	0.153	0.435	0.17	0.944	0.534
	False Positive	1.512	0.663	0.017	0.592	0.638	0.083	0.276	0.14	0.659	0.465
	False Negative	1.733	0.750	0.018	0.778	0.716	0.137	0.366	0.17	0.814	0.516
	P-values	< 0.001	< 0.001	>0.05	< 0.001	< 0.001	< 0.001	< 0.001	>0.05	< 0.001	< 0.001
BIO_CLINICALBERT	True Positive	1.374	0.697	0.015	0.623	0.675	0.097	0.307	0.14	0.688	1.374
	True Negative	1.993	0.784	0.02	0.734	0.752	0.130	0.369	0.14	0.802	1.993
	False Positive	1.705	0.784	0.014	0.725	0.706	0.111	0.345	0.17	0.781	1.705
	False Negative	1.806	0.733	0.022	0.747	0.698	0.123	0.329	0.14	0.784	1.806
	P-values	< 0.001	< 0.001	>0.05	< 0.001	< 0.001	< 0.001	< 0.001	>0.05	< 0.001	< 0.001