



Insights from Sentiment Analysis of Drug Reviews

Introduction

In this study, we are investigating the application of text mining in the pharmaceutical field to conduct sentiment analysis on online drug reviews. Analyzing user-generated content is essential for understanding their opinions regarding medication effectiveness and side effects, which can be used to enhance pharmacovigilance systems. The purpose of our experiments is to develop an automatic system that can effectively process a large volume of reviews and provide more objective insights. From a text-mining perspective, our experiments are interesting as they involve extracting meaningful information from unstructured data, dealing with challenges such as variations in language usage, and detecting sentiment polarities accurately.

This sentiment analysis on online drug reviews intends to analyze the language used in reviews and determine the overall sentiment towards a particular drug. By harnessing the power of sentiment analysis, it is possible to gain objective insights into the effectiveness and safety of drugs, which can in turn help to improve pharmacovigilance systems. Through this process, potential adverse effects can be identified and addressed more quickly, improving patient safety and overall health outcomes. By delivering objective insights, this analysis can help to inform regulatory decisions and promote more informed and effective healthcare practices.

We detail our implementation process for conducting sentiment analysis on patient reviews related to specific drugs and conditions. First, we gathered and preprocessed the data by downloading a dataset from the UC Irvine Machine Learning Repository, which provided patient reviews and 10-star ratings reflecting overall satisfaction. Next, we applied text mining techniques using LightSIDE and machine learning algorithms to extract meaningful information from the unstructured data. Through this process, we were able to analyze sentiment by identifying and measuring the polarity of sentiment in the reviews. This allowed us to determine users' opinions on various medications and gain insights into their effectiveness and safety. By leveraging these techniques, we were able to deliver objective insights to enhance pharmacovigilance systems and promote better healthcare practices.

Our project aims to contribute to the field of pharmaceutical landscape by gathering and analyzing users' opinions and experiences regarding medication quality and safety. By understanding these perspectives, we hope to inform decision-making processes and drive improvements in pharmacovigilance systems. Our ultimate goal is to enhance the detection, assessment, and prevention of adverse drug effects and related problems. Through this analysis, we aim to discover previously unidentified risks, side effects, or drug interactions to enhance patient protection and well-being. By utilizing text mining techniques, we can efficiently and effectively analyze vast amounts of patient feedback on different types of medication, ultimately contributing to the improvement of medication safety and efficacy. Through this whole semester, we have been exploring the field of text mining. This

technology enables researchers as well as us to analyze vast amounts of text and extract patterns, trends, and relationships that would be nearly impossible to discern manually. In order to help our analysis more comprehensively, we explored the following related works.

Related Work

Several studies have demonstrated the potential of natural language processing (NLP) and machine learning algorithms to extract valuable insights from unstructured textual data. Approaches include machine learning techniques (e.g., Naïve Bayes or Support Vector Machines), deep learning methods (e.g., Recurrent Neural Networks or Transformer models), and lexicon-based approaches. Most studies focus on either classifying overall sentiment or identifying the sentiment associated with specific aspects (e.g., effectiveness or side effects). Moreover, some researchers have utilized domain-specific knowledge by integrating medical ontologies or curated vocabularies. Our chosen approach aims to create a hybrid model combining both machine learning and deep learning techniques to address the limitations of each method individually. We will compare its performance with other methods in terms of accuracy, recall, precision, and F1-score. Additionally, we plan to enhance our approach by incorporating domain-specific knowledge in order to improve the granularity and relevance of the extracted insights. The following datasets, formula and algorithms helped us conduct this research for reference.

Data	#Train	#Test	#conditions	#drugs	length	rating	label	%
Drugs.com								
Overall Rating	161297	53766	836	3654	458.32 (240.76)	$rating \leq 4$	-1	25
						$4 < rating < 7$	0	9
						$rating \geq 7$	1	66
Side Effects (Annotated)	-	400	141	243	500.385 (209.42)	No Side Effects	0	32
						Mild / Moderate Side Effects	1	28
						Severe / Extremely Severe Side Effects	2	40
Druglib.com								
Overall Rating	3107	1036	1808	541	277.57 (283.21)	$rating \leq 4$	-1	21
						$4 < rating < 7$	0	10
						$rating \geq 7$	1	69
Benefits (Effectiveness)	3107	1036	1808	541	212.87 (198.51)	Ineffective	0	8
						Marginally / Moderately Effective	1	19
						Considerably / Highly Effective	2	73
Side Effects	3107	1036	1808	541	177.36 (197.93)	No Side Effects	0	30
						Mild / Moderate Side Effects	1	53
						Severe / Extremely Severe Side Effects	2	17

Figure 1 Data Description of Dataset in Related Work

		Train Data					avg. test
		Contraception	Depression	Pain	Anxiety	Diabetes, Type 2	
Test Data	Contraception	95.57 / 92.39	64.40 / 35.66	59.36 / 22.59	60.59 / 24.59	62.12 / 33.63	68.41 / 41.77
	Depression	62.05 / 31.51	90.13 / 78.07	75.21 / 40.69	77.07 / 43.95	66.98 / 33.93	74.29 / 45.63
	Pain	66.53 / 27.11	78.80 / 42.43	92.65 / 79.32	80.72 / 37.50	57.70 / 20.67	75.28 / 41.40
	Anxiety	64.35 / 28.14	82.64 / 51.22	79.74 / 43.43	92.37 / 78.41	67.51 / 30.64	77.32 / 46.37
	Diabetes, Type 2	69.90 / 44.50	71.83 / 43.37	68.17 / 32.32	69.48 / 34.18	94.74 / 89.84	74.82 / 48.84
	avg. train	71.68 / 44.73	77.56 / 50.15	75.03 / 43.67	76.05 / 43.73	69.81 / 41.74	

Figure 2 Cross-Domain Sentiment Analysis of Related Work

Aspect	Train Source	Test Source	Acc. / Kappa
Overall Rating	Drugs.com	Druglib.com	75.29 / 48.08
Overall Rating (all)	Druglib.com	Drugs.com	70.06 / 26.76
Side Effects	Druglib.com	Drugs.com	49.75 / 25.88

Figure 3 Cross-data Sentiment Analysis of Related Work

The author performs sentiment analysis to predict the sentiments concerning overall satisfaction, side effects and effectiveness of user reviews on specific drugs.

- Hossain, M. D., Azam, M. S., Ali, M. J., & Sabit, H. (2020). Drugs rating generation and recommendation from sentiment analysis of drug reviews using machine learning. We will refer to this paper and compare the algorithms and all of them have different performances concerning the prediction accuracy. It would help us explore how to design and implement a drug recommender system framework that applies sentiment analysis technologies on drug review.

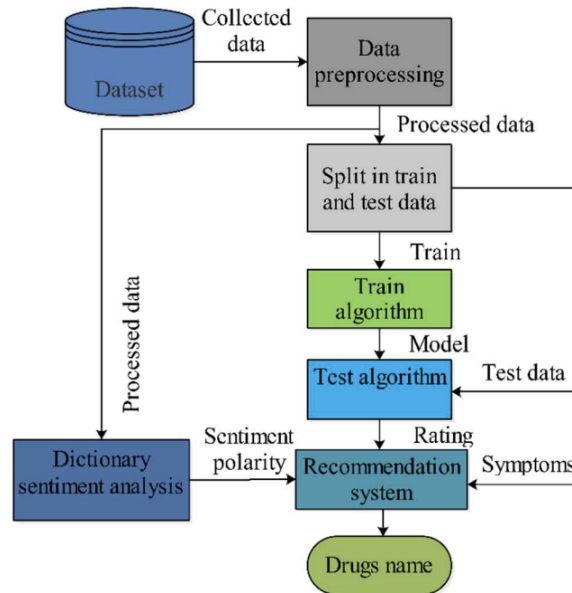


Figure 4 Framework of Drug Recommender System

$$f(x) = B_0 + \sum a_i(x, x_i)$$

$$f(x) = B_0 + \sum (x * x_i)$$

(The researchers used the Feature Extraction referring to Unigrams, Bigrams & Trigrams, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) which are similar to what we worked with LightSIDE.)

Lexicon-based approach

$$p(r) = \begin{cases} 1 & \text{if } \# \text{ positive} \geq \# \text{ negative} \\ -1 & \text{if } \# \text{ positive} < \# \text{ negative} \end{cases}$$

In order to evaluate the different approaches, we have computed the traditional measures in text classification: Precision (P), Recall(R), F-score (F1) and Accuracy (Acc).

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2PR}{P + R}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

(This research applied supervised learning and lexicon-based sentiment analysis approaches over two different corporas extracted from social platforms.)

Data Processing

The dataset we used is cited from provides patient reviews on specific drugs along with related conditions and a 10 star patient rating reflecting overall patient satisfaction. The data was obtained by crawling online pharmaceutical review sites. The data is split into a train (75%) a test (25%) partition (see publication) and stored in two .tsv (tab-separated-values) files, respectively. There are 215063 instances in total.

To implement our experiment, we wrote a python script to classify the conditions of reviews and save them as separate tables. We also removed reviews with ratings from 4 to 7 (inclusive) to enhance the features of terms. Then classified the reviews with positive reviews that have a rate higher than 5 and negative reviews that have a rate lower than 5.

In case of overemphasizing positive and negative features, we balanced the positive and negative proportions in all training sets with removing redundant data and duplicating the data.

Approaches & Experiment

Our research focuses on developing and evaluating sentiment analysis models for pharmaceutical reviews by exploring the applicability of our models to different conditions

and understanding the impact of positive and negative review proportions on prediction accuracy. We hope to gain valuable insights that can contribute to the broader field of sentiment analysis in the pharmaceutical domain.

Experiment 1

1.1 One Condition to All Reviews

In this experiment, we want to find whether the sentiment analysis model trained on drug reviews for a specific condition could be generalized to perform sentiment analysis on drug reviews for other conditions. The reason why we designed this experiment is based on the need for sentiment analysis models to be adaptable to different conditions and medications. This is because users may use different words and phrases to describe the same side effects, symptoms, or benefits of a medication for different conditions. Therefore, we hypothesized that the sentiment analysis model trained on drug reviews for a specific condition would have a good level of generalizability and could perform sentiment analysis on drug reviews for other conditions with reasonable accuracy.

We trained a sentiment analysis model on drug reviews with the drug reviews of Birth Control, and evaluated the model's performance on drug reviews of the drugs used to treat the same condition. The output from LightSide are as follows:

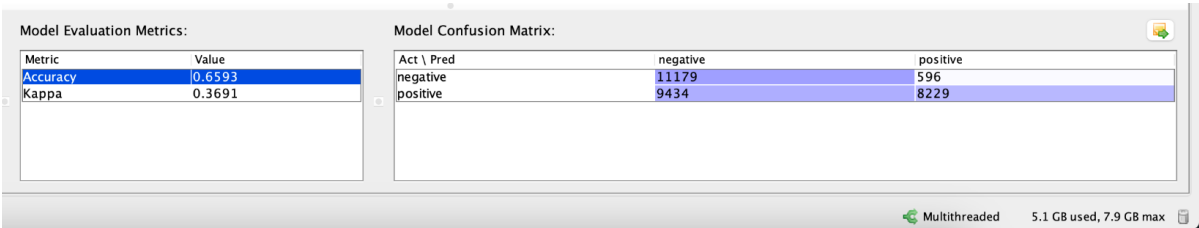


Figure 5 Output in Experiment 1.1 when pos/neg = 1.5

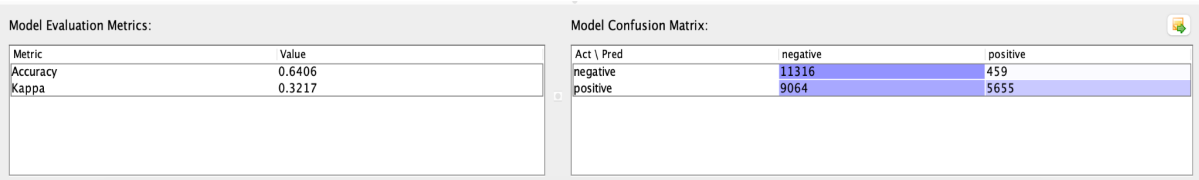


Figure 6 Output in Experiment 1.1 when pos/neg = 1.25

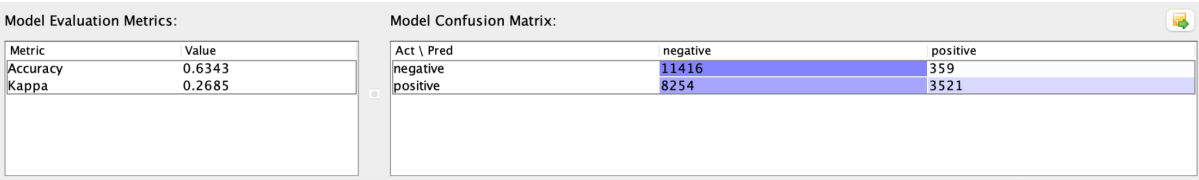


Figure 7 Output in Experiment 1.1 whe pos/neg = 1

Ratio of Positive to Negative reviews in Train Set(Pos/Neg)	Accuracy
---	----------

1.5	0.6593
1.25	0.6343
1	0.6406

Table 1 Result of experiment 1.1 and 1.2

It is not hard to find that the accuracy of every experiment is close which means the accuracy of predicting the sentiment of the drug views of the specific condition is limited.

1.2 Influence of an Unbalance Train Set

We thought that drugs that have been approved and are on the market have undergone rigorous scrutiny and testing by regulatory bodies, the testing and trials have already established the basic efficacy of these drugs in treating specific conditions. Therefore, we hypothesized that a higher proportion of positive reviews in the training or testing dataset for drug review sentiment analysis would improve the accuracy of the sentiment analysis model's predictions.

We control the ratio of the amount of positive reviews to the amount of negative reviews into 1.5, 1.25 and 1. So that we could find whether a higher proportion of positive reviews in the training set for drug review sentiment analysis would improve the accuracy of the model's sentiment analysis predictions.

As the result of the experiment presented in table 1, we can find that the results indicated that the sentiment analysis model's performance was not significantly affected by the proportion of positive reviews in the training set. This finding suggests that a balanced proportion of positive and negative reviews in the training set may be sufficient for training an accurate sentiment analysis model for drug reviews.

Experiment 2

This experiment aimed to validate the accuracy of sentiment analysis predictions on drug reviews for a specific condition. Specifically, we aimed to verify whether the sentiment analysis model trained on drug reviews for a specific medication and condition could be applied to perform sentiment analysis on drug reviews for other medications used to treat the same condition.

Every model was trained with Basic features, such as unigrams, bigrams, trigrams, pos bigrams, pos trigrams but without individual punctuation. Then, these models were built by Naive Bayes.

As the results of experiment 2 are presented in table 2, the accuracy of prediction is not higher due to the train sets and the test sets are based on the reviews of the same drugs. This result suggests that the prediction model which is trained by the reviews of a specific drug could be used to predict the drug reviews of the drugs for the same condition since they contains the same features of terms to describe a side effect or a phenomenon of recovering from this condition.



train\test	Etonogestrel	Ethinyl Estradiol / Norethindrone	Levonorgestrel	Nexplanon	Ethinyl estradiol / Levonorgestrel
Etonogestrel	0.8979	0.9038	0.8976	0.9401	0.8512
Ethinyl Estradiol / Norethindrone	0.858	0.9178	0.7413	0.8633	0.8705
Levonorgestrel	0.609	0.528	0.7708	0.5562	0.5537
Nexplanon	0.9278	0.8899	0.8993	0.8989	0.8375
Ethinyl Estradiol / Levonorgestrel	0.8406	0.9073	0.7188	0.8596	0.8457

Table 2 The result of experiment 2

However, the accuracy of prediction being tested by a model that was trained by the reviews of Levonorgestrel is apparently lower.

Why is the accuracy of prediction using a model that was trained by the reviews of Levonorgestrel apparently low?

- Strong subjectivity: Due to the fact that Levonorgestrel is an emergency contraceptive, users' usage scenarios and psychological states may differ from those of other drugs. This may make users more susceptible to emotional influences when evaluating the drug, leading to a decrease in the accuracy of sentiment analysis.
- Drug side effects: Levonorgestrel may cause some side effects such as headache, nausea, and fatigue. Different individuals may have varying perceptions and descriptions of the side effects, which may decrease the accuracy of sentiment analysis results.

Conclusion:

Through the two experiments, we find that the sentiment analysis model's performance was not significantly affected by the proportion of positive reviews in the training set. This finding suggests that a balanced proportion of positive and negative reviews in the training set may be sufficient for training an accurate sentiment analysis model for drug reviews. However, the possibility of a relationship between an extremely high or low proportion of positive reviews in the training set with the accuracy of drug sentiment analysis and prediction still exists. Also, the prediction model which is trained by the reviews of a specific drug could be used to predict the drug reviews of the drugs for the same condition since they contain the same features of terms to describe a side effect or a phenomenon of recovering from this condition. Moreover, the influence of the drugs on patients need to be considered in sentiment analysis of drug reviews.

References

Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). *Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning*. Proceedings of the 2018 International Conference on Digital Health.

Hossain, M. D., Azam, M. S., Ali, M. J., & Sabit, H. (2020). *Drugs rating generation and recommendation from sentiment analysis of drug reviews using machine learning*. 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE). <https://doi.org/10.1109/etcce51779.2020.9350868>

Cristóbal Colón-Ruiz, *Highlights•Drug reviews provide useful information to improve pharmacovigilance systems. •Sentiment analysis of drug reviews by hybrid deep learning classifiers. •BERT followed by bidirectional LSTM provides slightly better performance. •CNN achieves accepta, & Abstract*Since the turn of the century. (2020, August 17). *Comparing deep learning architectures for sentiment analysis on drug reviews*. Journal of Biomedical Informatics.

S. Garg, "Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 175-181, doi: 10.1109/Confluence51648.2021.9377188.

Vijayaraghavan, S., & Basu, D. (2020, March 21). *Sentiment Analysis in drug reviews using supervised machine learning algorithms*. arXiv.org. Retrieved February 20, 2023, from <https://arxiv.org/abs/2003.11643>.

Zafra, S.M., Valdivia, M.T., Molina-González, M.D., & López, L.A. (2019). *How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for the medical domain*. Artificial intelligence in medicine, 93, 50-57 . This research applied supervised learning and lexicon-based sentiment analysis approaches over two different corpora extracted from social web.

Na, J., & Kyaing, W.Y. (2015). *Sentiment Analysis of User-Generated Content on Drug Review Websites*. Journal of Information Science Theory and Practice, 3, 6-23.

Addition:

The Outputs of 25 experiments in Experiment2:

(A: Etonogestrel, B: Ethinyl estradiol / norethindrone, C: Levonorgestrel, D: Nexplanon, E: Ethinyl Estradiol / Levonorgestrel)

AA :

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.8979	negative	319	25
Kappa	0.7939	positive	57	402

AB :

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.9038	negative	251	23
Kappa	0.8076	positive	32	266

AC :

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.8976	negative	128	15
Kappa	0.743	positive	44	389

AD :

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.9401	negative	251	6
Kappa	0.8803	positive	26	251

AE :

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.8512	negative	139	23
Kappa	0.7004	positive	31	170

BA:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.858	negative	306	38
Kappa	0.7141	positive	76	383

BB:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.9178	negative	257	17
Kappa	0.8357	positive	30	268

BC:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.7413	negative	132	11
Kappa	0.4658	positive	138	295

BD:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8633
Kappa	0.7273

Model Confusion Matrix:

Act \ Pred	negative	positive
negative	235	22
positive	51	226

BE:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8705
Kappa	0.741

Model Confusion Matrix:

Act \ Pred	negative	positive
negative	148	14
positive	33	168

CA:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.609
Kappa	0.0985

Model Confusion Matrix:

Act \ Pred	negative	positive
negative	30	314
positive	0	459

CB:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.528
Kappa	0.0152

Model Confusion Matrix:

Act \ Pred	negative	positive
negative	4	270
positive	0	298

CC:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.7708
Kappa	0.1113

Model Confusion Matrix:

Act \ Pred	negative	positive
negative	11	132
positive	0	433

CD:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.5562	negative	21	236
Kappa	0.0808	positive	1	276

CE:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.5537	negative	0	162
Kappa	0	positive	0	201

DA:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.9278	negative	328	16
Kappa	0.8539	positive	42	417

DB:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.8899	negative	243	31
Kappa	0.7794	positive	32	266

DC:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8993
Kappa	0.7446

Model Confusion Matrix:


Act \ Pred	negative	positive
negative	126	17
positive	41	392

DD:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8989
Kappa	0.798

Model Confusion Matrix:



Act \ Pred	negative	positive
negative	240	17
positive	37	240

DE:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8375
Kappa	0.6725

Model Confusion Matrix:

Act \ Pred	negative	positive
negative	136	26
positive	33	168

EA:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.8406
Kappa	0.681

Model Confusion Matrix:

Act \ Pred	negative	positive
negative	308	36
positive	92	367

EB:

Model Evaluation Metrics:

Metric	Value
Accuracy	0.9073
Kappa	0.8147

Model Confusion Matrix:

Act \ Pred	negative	positive
negative	254	20
positive	33	265

EC:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.7188	negative	134	9
Kappa	0.4365	positive	153	280

ED:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.8596	negative	238	19
Kappa	0.7202	positive	56	221

EE:

Model Evaluation Metrics:		Model Confusion Matrix:		
Metric	Value	Act \ Pred	negative	positive
Accuracy	0.8457	negative	149	13
Kappa	0.6934	positive	43	158