Enhancing Prediction of Sentiment score for Patient Drug Review

Dipesh Yadav (36509669)
Lancaster University
School of Computing and Communication
SCC-413: Applied Data Mining
d.yadav1@lancaster.ac.uk

Abstract

This report details a comprehensive sentiment analysis aimed at understanding patient sentiments across various medications and conditions by analyzing drug reviews. Utilizing the dictionary, along with correct pre-processing path and robust model, the project provides evaluation of sentiments based on domain specific recommendation. The analysis integrates dictionary-based sentiment classification with predictive modeling to create a composite sentiment score, factoring in the utility of reviews for recommendation.

1 Introduction

1.1 Problem Definition

In the pharmaceutical industry, the effectiveness of drugs is paramount. Evaluating patient satisfaction across various illnesses through their feedback consumes time via manual process and accuracy is very crucial for recommendation systems. Accurately translating review sentiments requires careful selection of pre-processing methods and a strong model to reduce bias in sentiment analysis.

1.2 Motivation

The drug effectiveness and patient satisfaction is crucial in medical sector. This information assists in refining personalized treatment plans and identifying optimal drug alternatives, thus enhancing health outcomes. Increasing sentiment prediction of model ensures accurate, unbiased analysis with reliable, customized recommended treatment.

1.3 Research Topic

The research seeks to pinpoint "What is the effectiveness of drugs for a given condition and which factor improve the efficacy of sentiment based recommendation systems in the healthcare industry?"

1.4 Contribution

The research makes the following contributions:

a. It enhances sentiment analysis in pharmaceutical reviews by integrating the different emotional dictionary with machine learning models to improve sentiment score.

b. The pre-processing factors influencing model training and factors to consider during model selection.

2 Related Work

Research on sentiment analysis in healthcare sectors demonstrates complexities and effectiveness of different models. Study (Asp) introduces an automated method using BERT for aspect extraction and classification from reviews, showing superior efficiency over traditional methods. Another study (Wang et al., 2023) uses the UIMF method combining BERT with traditional machine learning to assess mental health drug treatment satisfaction, effectively balancing accuracy and speed while improving treatment outcomes.

Study (Alaie et al., 2023) focuses on sentiment classification of pharmaceutical reviews, finding Count Vector (CV) more effective than TF-IDF and SVC superior due to its margin classification, while Random Forest Classifier under performed due to limited feature incorporation.

(Gräßer et al., 2018) study applies machine learning to analyze sentiments in patient-generated drug reviews, focusing on satisfaction and effectiveness. Challenges suggesting deep learning and larger datasets could improve clinical applications.

Therefore, selecting appropriate machine learning techniques, datasets and data processing methods enhance sentiment analysis in healthcare, emphasizing the potential of advanced, automated, and hybrid models to address challenges of large, complex datasets for better decision-making.

3 Data

3.1 Data Definition

The study utilizes dataset Drug Reviews[1] (UCI) from the UCI Repository with 215,063 reviews, featuring six evaluative metrics and sentiments via reviews and ratings, split into train and test files.This data was published in a study on sentiment analysis of drug experience over multiple facets, ex. sentiments learned on specific aspects such as effectiveness and side effects. The Harvard IV-4 Psychosociological Dictionary[2] (DUNPHY, 1974), developed for the Harvard General Inquirer, aids in analyzing psychological and sociological content in texts for various research applications.

3.2 Data Sourcing

The review dataset is obtained by crawling online pharmaceutical review sites and is accessible publicly under a CC BY 4.0 license via the UCI Repository. This ensures it can be utilized for academic and research purposes with appropriate citation. The Harvard-IV data is sourced from (DUNPHY, 1974). The dataset is available for general emotional analysis for research.

3.3 Rationale for Appropriateness

This dataset is ideal for our research as it centers on the pharmaceutical industry. Its substantial size enables unbiased analysis and facilitates the evaluation of different pre-processing methods and model selections. Additionally, the inclusion of an emotional dictionary enhances sentiment analysis effectiveness.

3.4 Limitations

One limitation of the dataset is its reliance on reviews from a single source, which may not capture the full spectrum of patient experiences and sentiments. Additionally, the dataset's focus on pharmaceuticals may limit its generalization to other industries or healthcare domains. Moreover, there may be inherent biases in the reviews, such as selection bias or varying levels of detail, which could impact the analysis results.

4 Methodology

The research is divided in four sections Exploratory Data Analysis, Pre-processing, Model Selection and Model Prediction.

[1]https://doi.org/10.24432/C5SK5S.

[2]https://inquirer.sites.fas.harvard.edu/homecat.htm

N-Gram	Accuracy(%)	Loss Metrics
3-Gram	61.2	1.558
4-Gram	64.5	1.105

Table 1: Deep Learning N-Gram based model accuracy.

4.1 Exploratory Data Analysis

In analyzing drug review data, initial filtering identified conditions with unique, non-contributing drugs and anomalies in data for certain conditions. The word cloud helped in revealing commonly used stop words Figure 3. Reviews were categorized by ratings for targeted sentiment analysis (Cao et al., 2021).

While uni-gram, bi-gram, and 3-gram showed similar patterns across positive and negative reviews, 4-gram effectively differentiated between the two. Trends over time and average ratings per year were tracked to observe shifts in user engagement and sentiment evolution (Dey et al., 2018).

Analysis of the 'usefulCount' metric, which varied significantly Figure 2, indicated varying engagement levels and the influence of drug popularity on review utility. Given the low percentage of missing data, data removal were considered necessary. 3-gram model accuracy is found less than 4-gram model as per Table 1 suggesting nuanced language usage in more detailed phrases, highlighting the challenge of over-fitting in deep learning models (Lau et al., 2023).

4.2 Pre-Processing

Handling missing data ensures a clean dataset, crucial for analysis. Removing HTML tags and other non-text elements is essential, especially with dataset prepared by scraping from Drugs.com.

For a recommendation system, analyzing conditions with only one drug isn't feasible. Therefore, the analysis will focus on conditions that have at least two drugs to avoid bias(Con).

Standard English stop-words from NLTK are used to filter non-essential words in text analysis. Additionally, a list including negations and sentiment-critical words is created using Word Cloud and excluded from stop-words to preserve crucial meanings and maintain sentiment accuracy in text processing tasks.

The pre-processing of review data utilizes libraries like scikit-learn, TensorFlow, and Keras for various machine learning tasks. This process includes using BeautifulSoup to remove HTML tags

Model	Accuracy(%)
Deep Learning(Keras)	64.5
Lightgbm(Rating Based)	70.0
Lightgbm(TextBlob Based)	84.5

Table 2: Model Accuracy.

from web-scraped data, applying regular expressions to filter out non-letter characters, and converting texts to lowercase for standardization. Additionally, the refined list of filtered stop-words is used to exclude non-essential but sentiment-critical words. The final step involves stemming with NLTK's SnowballStemmer to simplify words to their root forms, streamlining the text for more effective analysis (Khanal et al., 2020).

4.3 Model Selection

In sentiment analysis, ratings above 5 are labeled positive (1) and 5 or below as negative (0), simplifying it to a binary task. A CountVectorizer configured for 4-grams and limited to 20,000 features converts text to numerical data, focusing on richer context while managing feature dimensionality. A pipeline including the vectorizer ensures consistent pre-processing across training and testing datasets (N. and K.m., 2023).

The overfitting in a Deep Learning Keras model is observed due to significantly higher training accuracy compared to test accuracy and a substantial increase in loss metrics. Consequently, we switched to a LightGBM classifier, which offers higher accuracy for classification tasks as in Table 2. LightGBM is preferred for its efficiency with large datasets, as well as its speed and lower memory usage (Khanal et al., 2020).

4.4 Model Prediction

For sentiment analysis data preparation we use the TextBlob library, which assigns a sentiment score to reviews. These scores range from -1 (very negative) to 1 (very positive) and are added as new columns to the dataset, labeled "Predict Sentiment" and "Predictive Sentiment 2" for clean and original data respectively. Incorporating these sentiment scores enhances the dataset, providing valuable insights for predictive analytics (Aljedaani et al., 2022).

To enhance predictive sentiment analysis, new features based on review characteristics are integrated into the training model. These features include sentence count, word count, unique word count, letter count, punctuation count, uppercase and title case word counts, and stop-word count. Additionally, seasonal categorization based on certain period provide further insights into data patterns. The LightGBM classifier is strategically configured with specific hyper parameters such as boosting rounds, learning rate, tree depth, regularization, and sub-sampling. Early stopping is implemented to prevent over-fitting. The model uses 80% of the data for training and the remaining 20% for validation. Post-training, predictions and feature importance are evaluated to identify the most influential features, guiding future model adjustments and feature optimizations. This comprehensive approach ensures robust training and valuable insights for enhancing sentiment analysis accuracy (Rec).

Harward dictionary is used to increase the predictive sentiment score for pharmaceutical sector. The lists of positive and negative words are extracted from the Harvard emotional dictionary to precisely analyze sentiments in reviews. This setup uses the CountVectorizer to quantify sentiments by counting specific emotional words. Additionally, sentiment ratios are calculated to classify overall sentiment based on the prevalence of positive versus negative words.

Moreover, a function normalizes review usefulness by the frequency of conditions, providing contextually rich insights. Finally, a composite sentiment score is created by integrating predictions from multiple models, weighted by review usefulness, and aggregated to summarize sentiments for specific drugs under various conditions. This multifaceted approach leverages both linguistic and contextual data, enhancing predictive accuracy and providing actionable insights into drug effectiveness across different conditions.

5 Results and Findings

The TensorFlow model displayed improved training performance over 10 epochs, with accuracy increasing from 71.24% to 80.38%, and loss decreasing from 0.5831 to 0.3981, demonstrating effective learning. However, a testing accuracy of 64.51% suggested potential overfitting or a need for parameter tuning to better generalize to new data. refer Figure 4.

TextBlob sentiment analysis demonstrated a moderate correlation with numerical ratings (0.348) and binary sentiment (0.317), indicating that the

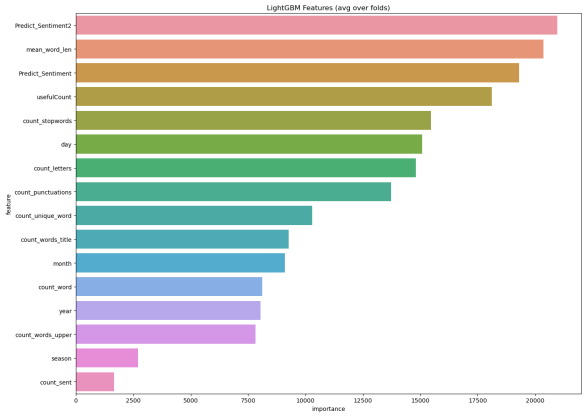

Figure 1: Feature Importance.

computed sentiment scores align with the general sentiment trends in ratings and classifications. Each review was successfully processed, receiving a sentiment score that accurately reflects a spectrum of sentiments from negative to positive, based on review content.

Conversely, the LightGBM model, employing early stopping after 100 rounds without improvement, excelled with a final iteration at 6768, showcasing a well-balanced confusion matrix True Positive 45341, True Negative 13767, False Positive 7242, False Negative 3628 and robust prediction capabilities across classes, achieving an accuracy of 84.5% for review based feature selection refer Figure 1.

The sentiment analysis framework further utilizes a Harward dictionary to categorize drug review sentiments into positive, negative, or neutral replacing package dictionary, with substantial feature engineering in text data. The model integrates multiple sentiment predictions weighted by review usefulness. This approach highlights varying drug efficacy across conditions, guiding further validation and model adjustments for improved healthcare outcomes due to domain specific emotional dictionary for sentiment prediction.

We find that the research integrated model comparison and used Harvard emotional dictionary to enhance domain based drug review sentiment analysis. LightGBM outperformed in accuracy, showing less overfitting. Tailored tools and extensive preprocessing provided accurate, domain-specific insights, improving drug recommendations and highlighting the potential of sophisticated sentiment analysis in healthcare.

6 Conclusion and Future Work

The research effectively combined TensorFlow, LightGBM, TextBlob, and a Harvard emotional dictionary to enhance sentiment analysis for drug reviews. LightGBM demonstrated higher accuracy and reduced overfitting compared to TensorFlow. The tailored use of TextBlob and the emotional dictionary provided more precise sentiment assessments, improving the quality of drug recommendations within the pharmaceutical context.

6.1 Implication

This research underscores the value of domain-specific sentiment analysis in healthcare, enhancing pharmaceutical evaluations and recommendations through advanced machine learning and tailored tools. Improved understanding of patient feedback can lead to better outcomes and personalized treatments. The findings encourage ongoing refinement and adaptation of models to ensure accurate, relevant insights, setting a benchmark for future sentiment analysis applications in various sectors.

6.2 Limitation

First, the reliance on the emotional dictionary for sentiment analysis can lead to biased results, especially in texts with few sentiment words, where a small number of words might disproportionately affect the sentiment classification. Also, the normalization of sentiment scores using usefulCount may introduce a temporal bias, favoring older reviews that have accumulated more useful counts over time, potentially skewing the analysis towards outdated sentiments. Lastly, the method does not account for sentiment polarity when weighting by usefulCount, possibly misrepresenting the true impact and reliability of expressed sentiments.

6.3 Future Work

Future enhancements for the project include expanding the sentiment dictionary, introducing dynamic thresholding, and incorporating time-weighted normalization for usefulCount. Adjustments for sentiment polarity, employing advanced machine learning and deep learning techniques, and establishing a continuous feedback loop for model updates are also planned. These improvements aim to refine the accuracy and applicability of sentiment analysis in healthcare.

References

Aspect extraction and classification for sentiment analysis in drug reviews - ProQuest.

The context-based review recommendation system in e-business platform - ProQuest.

Recommendation Medicines by using a review.

UCI ML Drug Review dataset.

Aaqib Iqbal Alaie, Umar Farooq, Wakeel Ahmad Bhat, Surinder Singh Khurana, and Parvinder Singh. 2023. An Empirical Study on Sentimental Drug Review Analysis Using Lexicon and Machine Learning-Based Techniques. *SN Computer Science*, 5(1):63.

Wajdi Aljedaani, Furqan Rustam, Mohamed Wiem Mkaouer, Abdullatif Ghallab, Vaibhav Rupapara, Patrick Bernard Washington, Ernesto Lee, and Imran Ashraf. 2022. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. *Knowledge-Based Systems*, 255:109780.

Hang Cao, E. Zeynep Erson-Omay, Murat Günel, Jennifer Moliterno, and Robert K. Fulbright. 2021. A Quantitative Assessment of Pre-Operative MRI Reports in Glioma Patients: Report Metrics and IDH Prediction Ability. *Frontiers in Oncology*, 10:600327.

Atanu Dey, Mamata Jenamani, and Jitesh J. Thakkar. 2018. *Senti-N-Gram*: An *n*-gram lexicon for sentiment analysis. *Expert Systems with Applications*, 103:92–105.

D.C. DUNPHY. 1974. *Harvard iv-4 dictionary general inquirer project*. UNIVERSITY OF NEW SOUTH WALES.

Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. In *Proceedings of the 2018 International Conference on Digital Health*, pages 121–125, Lyon France. ACM.

Shristi Shakya Khanal, P. W. C. Prasad, Abeer Alsadoon, and Angelika Maag. 2020. A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 25(4):2635–2665. Publisher: Springer.

Clinton Lau, Xiaodan Zhu, and Wai-Yip Chan. 2023. Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry*, 14:1160291.

Vedavathi N. and Anil Kumar K.m. 2023. E-learning course recommendation based on sentiment analysis using hybrid Elman similarity. *Knowledge-Based Systems*, 259:110086.

Yi Wang, Yide Yu, Yue Liu, Yan Ma, and Patrick Cheong-Iao Pang. 2023. Predicting Patients' Satisfaction With Mental Health Drug Treatment Using Their Reviews: Unified Interchangeable Model Fusion Approach. *JMIR Mental Health*, 10:e49894.

7 Appendix

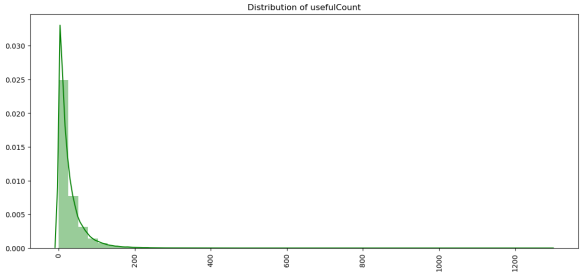

Figure 2: Distribution of usefulCount.

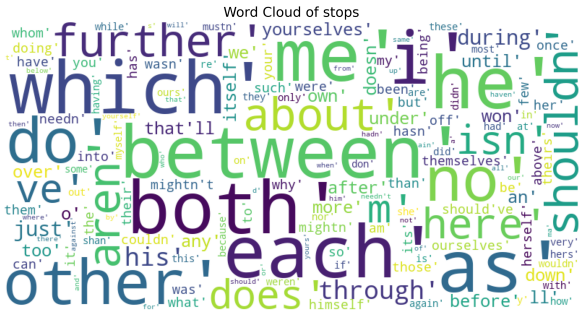

Figure 3: Word Cloud to determine Stop-words.

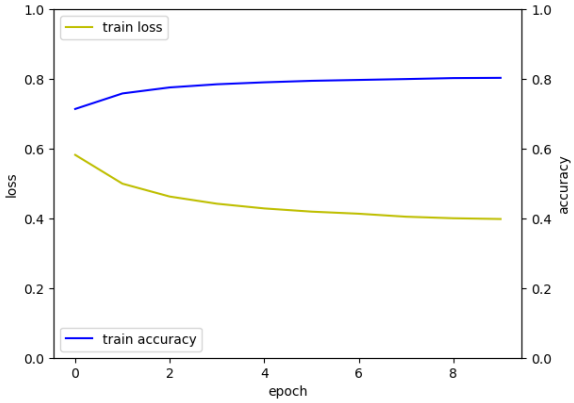

Figure 4: Epoch train loss and accuracy.