Enhancing Drug Reviews Insights through Exploratory Data Analysis and Sentiment Analysis

Ana Sofia Pinto*, Matilde Pato*†‡§, Nuno Datia*§
*Lisbon School of Engineering (ISEL) Politécnico de Lisboa, 1959-007 Lisbon, Portugal
‡Instituto de Biofísica e Engenharia Biomédica (IBEB), Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisbon, Portugal,
‡LASIGE, Faculdade de Ciências da Universidade de Lisboa, Campo Grande, 1749-016 Lisbon, Portugal
§NOVA LINCS, NOVA School of Science and Technology, Largo da Torre, 2829-516 Caparica, Portugal
Email: *a46506@alunos.isel.pt, †matilde.pato@isel.pt

2024 28th International Conference Information Visualisation (IV) | 979-8-3503-8016-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IV64223.2024.00042

***Abstract*—The increasing volume of user-generated content across various online platforms has created vast datasets in multiple domains, including healthcare. This article explores the significant roles of data visualisation and sentiment analysis within the healthcare sector using the UCI ML Drug Review dataset. Our study highlights the value of exploratory data analysis and sentiment analysis in comprehending patient feedback, enriching insights from the dataset. Data visualisation effectively elucidates the data's distribution and key characteristics, while sentiment analysis, performed using TextBlob and VADER, categorises the emotional tone of patient reviews. Our methodology aims to provide a deeper understanding of patient satisfaction and medication efficacy based on user-generated content.**

***Index Terms*—NLP, Sentiment Analysis, Exploratory Data Analysis, Drug Recommendation System, Healthcare Support**

I. Introduction

In pharmaceutical research, analysing patient reviews is crucial for assessing drug effectiveness and patient satisfaction. With the growth of online health forums and review platforms, it is essential to systematically process and analyse such data to enhance healthcare outcomes. Utilising the UCI ML Drug Review dataset [1], which contains extensive patient reviews, this study demonstrates how data visualisation and sentiment analysis can provide valuable insights in healthcare settings.

In this research, we explore into the intersection of healthcare and rising technology, showing the possibility to provide healthcare professionals and patients with valuable insights that can guide decision-making in treatment options, particularly in situations where numerous medications are available for a single condition, and individual medical histories vary widely. The need for such research has been emphasised by the COVID-19 pandemic, which severely strained healthcare systems and highlighted the importance of timely medical advice when direct healthcare access is constrained [2].

Our methodology consists of three phases: A. Exploratory Data Analysis, B. Cleaning and Pre-processing and C. Sentiment Analysis, for a better understanding of the contents of this paper here are the two main definition:

Exploratory Data Analysis (EDA) is the process of investigating datasets, elucidating subjects, and visualising outcomes. EDA is an approach to data analysis that applies a variety of techniques to maximise specific insights into a dataset, reveal an underlying structure, extract significant variables, detect outliers and anomalies, test assumptions, develop models, and determine best parameters for future estimations [3];

Sentiment Analysis (SA) also called *opinion mining*, is the field of study that analyses (with the use of Natural Language Processing – NLP) people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organisations, individuals, issues, events, topics, and their attributes [4]. In the context of this work, SA is the use of NLP to identify the sentiment behind patient reviews about a specific drug. By using the SA tools TextBlob and VADER, we analyse the emotional tone and content of patient reviews to categorise sentiments and discern patterns that can indicate medication efficacy and patient satisfaction.

This comprehensive approach not only aids in understanding patient feedback but also supports a more informed patient community. Individuals tend to trust reviews from peers who have experienced similar medical conditions and offer insights without personal gain. By systematically analysing such reviews, our study contributes to a more nuanced understanding of patient experiences and outcomes, facilitating better-informed choices in healthcare. This research is part of a broader effort to integrate data-driven insights into healthcare practice, aiming to improve the accuracy and relevance of information that patients and professionals rely on to make critical health decisions, since many side effects for different patients can be found in reviews like these ones.

This paper is structured as follows: Section II presents the Related Work, Section III describes the Methodology followed by Section IV, Results and Discussion and Section V concludes the paper and presents future work directions.

This work is supported by IBEB Research Unit ref. UIDB/00645/2020 (https://doi.org/10.54499/UIDB/00645/2020), NOVA LINCS Research Unit, ref. UIDB/04516/2020 (https://doi.org/10.54499/UIDB/04516/2020), and the LASIGE Research Unit, ref. UIDB/00408/2020 (https://doi.org/10.54499/UIDB/00408/2020) and ref. UIDP/00408/2020 (https://doi.org/10.54499/UIDP/00408/2020) with financial support from FCT, through national funds.

2375-0138/24/$31.00 ©2024 IEEE
DOI 10.1109/IV64223.2024.00042

The code and dataset developed for this work is available at https://github.com/matpato/EDRISA.git.

II. Related Work

This section explores recent literature that has increasingly recognised the importance of data-driven approaches in many sectors, including healthcare. The authors of [5] showed the importance and the vast possibilities for EDA, and the SA was performed on a subset of 8 conditions where many drugs were suggested and there were a great number of reviews. With this subset the authors performed SA using TextBlob for text classification as positive, negative and neutral.

Authors in [6] perform SA of drug-related discourse on social media platforms such as Twitter, Reddit, and Quora. They study the public opinion on drug consumption by analysing posts from different income-level countries. The research indicates that high-income countries display a more positive sentiment towards drug use. In contrast, low-income countries show a declining trend in positive sentiment, suggesting a disapproval of drug consumption. Factors like socio-economic conditions, cultural norms, and drug legalisation influence these sentiments. This analysis can inform public health policies and educational campaigns on drug use.

In [7], the authors explore the nuanced domain of SA on Twitter, focusing particularly on the diverse methodologies available for processing and analysing public sentiment during the COVID-19 pandemic in the UK. They conduct a comprehensive comparison between lexicon-based and machine-learning-based sentiment analysis techniques, emphasising the effectiveness of these methods in interpreting public opinion during different lockdown stages. The study not only explores sentiment classification and polarity analysis but also discusses the pre-processing steps necessary for effective SA. The authors provide a critical look at the implications of SA results, suggesting how these insights could guide public health policies and communication strategies during pandemic responses.

In [8], the authors declare that '*With the increasing usage of drugs to remedy different diseases, drug safety has become crucial over the past few years*', '*Consequently, a system that can help in determining post-release drug safety is a task of great significance*'. The authors claimed that although the application of SA in several domains such as business, hotel industry, recommendation system, etc., its use for drug reviews remains yet an under explored area. They explored a method to analyse drug safety reviews by combining lexicon-based and deep learning techniques. Using a dataset from 'Drugs.com', which includes reviews on drug-related side effects and reactions, firstly by applying Textblob to determine whether the sentiments in these reviews are positive, negative, or neutral. Then, for a more nuanced analysis, they employed a novel hybrid deep learning model that integrates Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks. This approach aimed to effectively classify the reviews, providing valuable insights into patient experiences with various drugs.

Complementing the SA in healthcare, the authors in [9] contributed to smart city analytics by assessing social media sentiments on urban services. Their innovative use of lexicon-based analysis and machine learning to interpret public feedback from smart city apps achieved a notable accuracy of 84%, underscoring the versatility of NLP tools in varied contexts. Their methodology informs our approach by showcasing the effectiveness of SA in enhancing service delivery within technologically-integrated city infrastructures.

III. Methodology

This section outlines the entire process of preparing and analysing the dataset. It is organised into three main sections: *A. Data Description*, which details the data collection methods and characteristics; *B. Data Cleaning and Pre-processing*, focusing on the preparation of the data; and *C. Sentiment Analysis*, which involves the analysis of sentiment within the data.

A. Dataset Description

The UCI Drug Review dataset [1], compiles patient reviews of various drugs alongside associated medical conditions and a 10-star rating system that reflects overall patient satisfaction. The dataset, derived from crawling online pharmaceutical review sites, encompasses over 200.000 patient reviews formatted as .csv files. The data is split into two partitions: `drugsComTest_raw.csv` for testing and `drugsComTrain_raw.csv` for training, which are 27,64 MB and 82,99 MB respectively, totalling 110,63 MB. The dataset comprises seven attributes: (1) `uniqueID` – a unique identifier for each review, (2) `drugName` – the name of the drug, (3) `condition` – the medical condition being treated, (4) `review` – the free text of the review, (5) `rating` – a patient-provided rating on a 10-point scale, (6) `date` – the date the review was entered, and (7) `UsefulCount` – the number of users who found the review useful.

B. Exploratory Data Analysis

In our Exploratory Data Analysis (EDA), we focused on thoroughly examining the dataset to understand its structure and contents fully. The EDA process involved reviewing the dataset attributes and assessing their distribution to gain insights into the dataset's composition. This phase is pivotal not only for validating the data format and content but also for evaluating the effectiveness of data cleaning and pre-processing efforts and guiding the subsequent SA discussions and potential future endeavours.

C. Data Cleaning and Pre-processing

To ensure the integrity and quality of our analysis, the dataset undergoes cleaning and pre-processing. For this, all the attributes were checked for missing values and other possible inconsistencies. Also, the removal of the attribute `rating` will take place in this phase, so future ranking of features is only based on SA. Then the focus will be on the steps present on the NLP pipeline, utilising the NLTK toolkit for

the individual pre-processing of the attribute `review`, going from raw text to cleaned text – the NLP pipeline steps were based on the observed needs of the reviews present in the dataset.

D. Sentiment Analysis

For sentiment classification we utilised the lexicon-based tools TextBlob and VADER – from NLTK. Reviews are classified into negative, neutral, and positive categories, and sentiments are rescaled to fall within a 0 to 10 range, facilitating a detailed comparison against conventional rating systems. With TextBlob and VADER, we analyse sentiments in both cleaned and raw reviews formats. The goal is to accurately quantify and understand the expressed sentiments, enhancing our interpretation of these critical data points, enriching the comprehension of patient feedback and contributes valuable insights into the dynamics of patient reviews in the healthcare sector.

IV. Results and Discussion

This section encloses the development of the *A. Exploratory Data Analysis*, the *B. Data Cleaning and Pre-processing,* mainly on the attribute *review*, and *C. Sentiment Analysis*, which involves classifying the polarity of the attribute review.

A. Exploratory Data Analysis

The dataset for the EDA has a size of (215.063, 7), and TABLE I shows the number of instances for the five different and relevant attributes.

Fig. 1 depicts the relation between the percentage of reviews and the rating they have. Drugs with rating 10 are the most reviewed, with a percentage of 31,62%, then 9 and 1 with 17,07% and 13,45%, respectively. Fig. 2 shows the relation between the percentage of the attribute `UsefulCount` and the rating they have. Drugs with rating 10 have the reviews considered most useful, with a percentage of 42,32%, then 9 and 8 with 20,60% and 12,18% respectively, and then 1, with a percentage of 7,53%. These visualisations are helpful for understanding how users generally feel about a product or service based on the reviews or `UsefulCount`.

Fig. 3 and Fig. 4 depict the top 10 drugs based on review count along their average ratings, and the top 10 drugs ranked by `UsefulCount` alongside their average ratings, respectively. The average for the top 10 most reviewed drugs is 6,67, whereas for those ranked by `UsefulCount` is 7,52. The top 10 drugs shown in the two figures are different, only sharing the drugs Sertraline and Escitalopram in different positions and Phetermine at number 7. With the intention to choose a reference of the 'Top 10 drugs' for future analysis, will give more importance to the attribute `UsefulCount` since it has much more data than the attribute reviews (since it has 'silent comments') and it has a bigger mean mean rating. Fig. 5 depicts the top 10 conditions by `UsefulCount`.

TABLE I: Number of instances for five attributes.

Attribute	Count
`drugName`	3671
`condition`	916
`UsefulCount`	602.1980
`uniqueID`	21.5063
`review`	21.5063

Fig. 6 and Fig. 7 show the number of unique conditions (`condition`) for the top 10 drugs by `UsefulCount` and the number of unique drugs (`drugName`) for the top 10 conditions by `usefulCount`. Analysing the bar graph, Fig. 6, it shows that from the top 10 drugs by `usefulCount`, Gabapentin includes the greater number of conditions, this number being 31, then Zoloft with 21 and Phentermine with the least number of unique conditions (`condition`), this being only 3. Fig. 7 depicts that Pain, Birth Control and High Blood Pressure are the Conditions with the greater number of unique drugs associated to them (from the top 10 conditions (`condition`) by `UsefulCount`) with 219, 181 and 146 unique drugs respectively. Weight Loss is the condition with the least number of unique drugs (`drugName`) only having 22 unique drugs associated.

Furthermore, we explored the top 10 prevalent `drug-condition` pairs for the top 10 conditions (`condition`) by `UsefulCount`, Fig. 5, so for each condition we analysed the top 10 drugs by `UsefulCount` and if this drugs were in the overall 'top 10 drugs by `UsefulCount`', Fig. 4. A summary of the drugs common to both the top 10 drugs of the `condition` by `UsefulCount` and the overall top 10 drugs by `UsefulCount` can be seen in TABLE II. It is important to know, that some drugs are common but simply are not presented in the top 10 drugs for the specific condition by `UsefulCount`.

The Fig. 8 depicts the usefulness vs. rating, the rating with the highest total useful count is 10 (Total useful count: 254.8464) and the rating with the lowest total useful count is 4 (Total useful count: 110241), and Fig. 9 depicts the usefulness vs. review length, the length with the most useful review having 187 words. The Fig. 9 also shows that the shortest reviews tend to be considered more useful. Finally, Fig. 10 presents the review sentiment over time (2008 until 2018), analysing closely the sentiment over the years we observed that in 2008 there was only 2% of negative comment, increasing to around 18% in the years on 2009 to 2012. In the years of

TABLE II: Prevalent drug-condition pairs

Condition	Common Drugs
Depression	Bupropion, Sertraline, Escitalopram, Citalopram, Duloxetine, Cymbalta
Anxiety	Escitalopram, Lexapro
Birth Control	None
Pain	Gabapentin
Bipolar Disorder	None
Weight Loss	Phentermine
Obesity	Phentermine
Insomnia	None
ADHD	None
High Blood Pressure	None

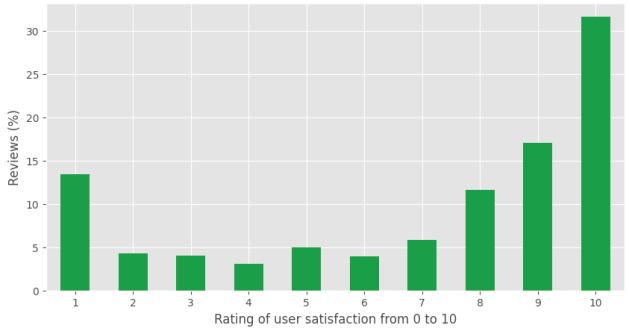

Fig. 1: Count of Reviews by Rating (%)

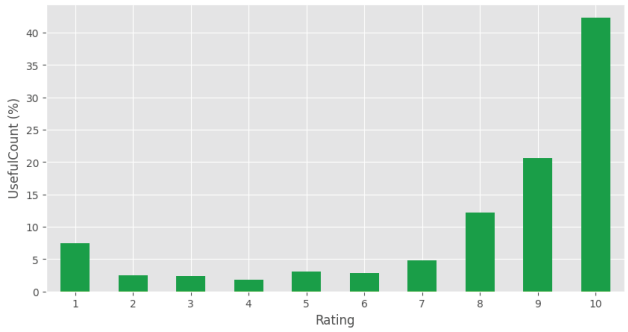

Fig. 2: UsefulCount by Rating (%)

2013 and 2014 there was a decrease of negative comments rounding the 12% and then an exponentially increased in the years 2015 to 2017, with the percentages, 25,8%, 34,9% and 36,9%, respectively.

B. Data Cleaning and Pre-processing

A thorough data cleaning and pre-processing phase was initiated to prepare the dataset for deeper analysis. Initially, we identified and removed corrupted rows based on inaccurate `condition` attribute. These 900 entries incorrectly contained user feedback counts instead of valid medical conditions, thus skewing the dataset's reliability. Following this, we enhanced the textual data through a pre-processing pipeline. This involved converting all text to lowercase, stripping `HTML` entities, and removing punctuation. We also filtered out stopwords and applied lemmatization to normalise the words, which can prove crucial for accurate text analysis. Furthermore, we removed the attribute `rating` and introduced the new column featuring the processed reviews. Finally, we saved the cleaned data into a .csv file (`_cleaned.csv`), ensuring it was well-structured and primed for the subsequent analysis. In the TABLE III we can see a specific review after the pre-processing pipeline.

In the cleaning process it was evaluated the possibility to remove drugs with a low number of reviews, but it was discarded since 32,42% of the dataset (train and test datasets) is composed of drugs with only one or two reviews.

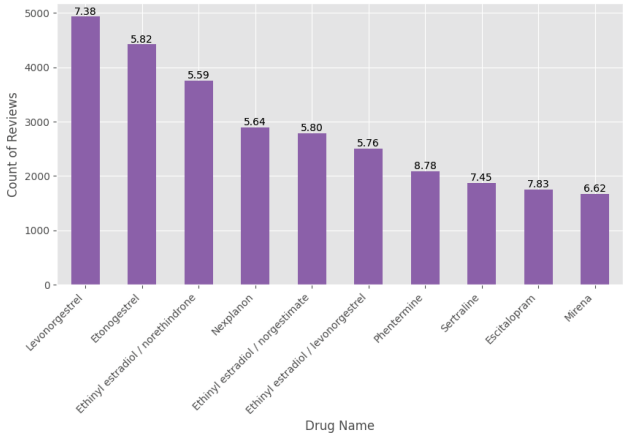

Fig. 3: Top 10 Most Reviewed Drugs with Their Mean Ratings

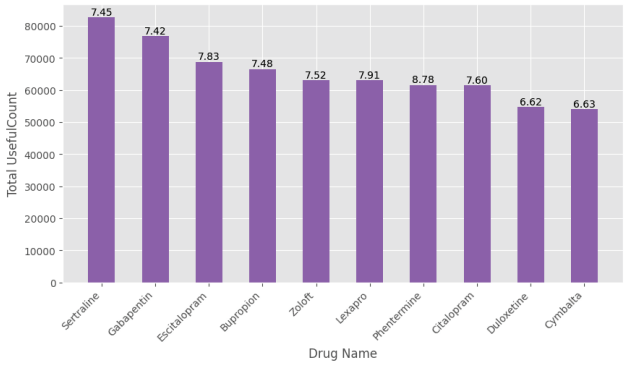

Fig. 4: Top 10 Drugs by UsefulCount and Their Mean Ratings

C. Sentiment Analysis

In the SA, firstly essential functions were defined for rescaling sentiment scores and labelling sentiment based on these

TABLE III: Comparison of Raw and Processed Review Texts

Type	Content
Raw Review	"*Abilify changed my life. There is hope. I was on Zoloft and Clonidine when I first started Abilify at the age of 15. Zoloft for depression and Clondine to manage my complete rage. My moods were out of control. I was depressed and hopeless one second and then mean, irrational, and full of rage the next. My Dr. prescribed me 2mg of Abilify and from that point on I feel like I have been cured though I know I'm not. Bipolar disorder is a constant battle. I know Abilify works for me because I have tried to get off it and lost complete control over my emotions. Went back on it and I was golden again. I am on 5mg 2x daily. I am now 21 and better than I have ever been in the past. Only side effect is I like to eat a lot.*"
Processed Review	"abilify change life hope zoloft clonidine first start abilify age 15 zoloft depression clondine manage complete rage mood control depress hopeless one second mean irrational full rage next dr prescribe 2mg abilify point feel like cure though know im bipolar disorder constant battle know abilify work try get lose complete control emotion go back golden 5mg 2x daily 21 good ever past side effect like eat lot"

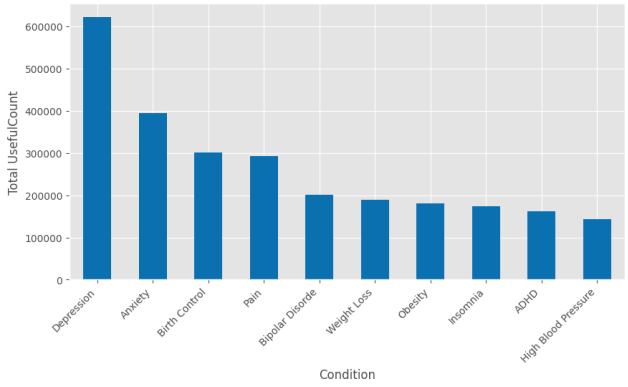

Fig. 5: Top 10 Conditions by UsefulCount

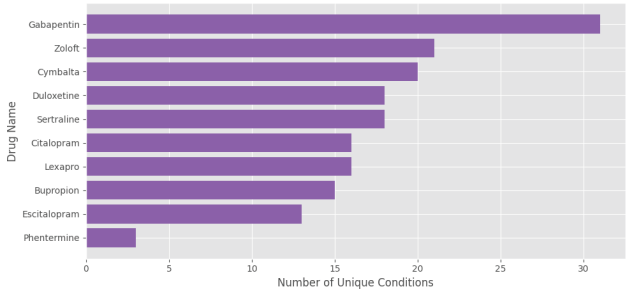

Fig. 6: Number of Unique Conditions for the Top 10 Drugs by UsefulCount

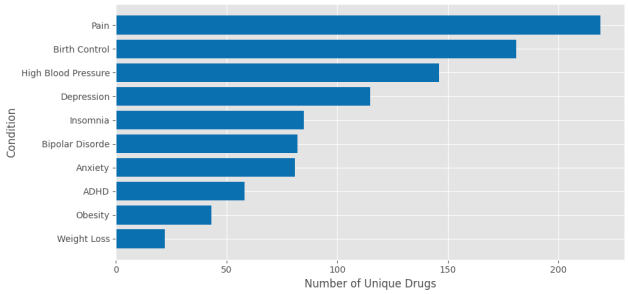

Fig. 7: Number of Unique Drugs for the Top 10 Conditions by UsefulCount

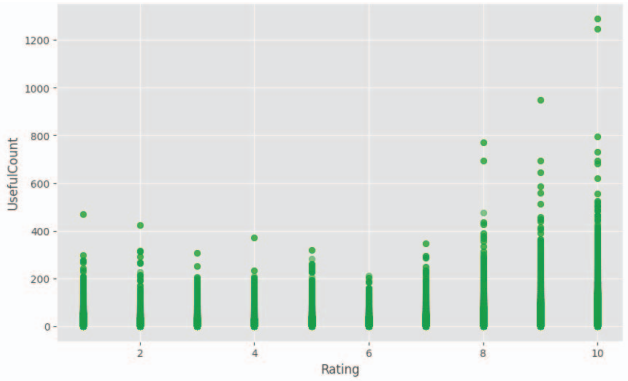

Fig. 8: Usefulness vs. Rating

rescaled scores. Following this, we implemented a function designed to process the dataset by conducting SA using both TextBlob and VADER tools. This function adjusts the sentiment scores, applies labels to them and ultimately saves the results into separate .csv files. We executed this function twice to accommodate both the raw and cleaned versions of the dataset. This workflow resulted in four output .csv files, capturing the SA results — from both TextBlob and VADER for each data type:

1) drugsComTrain_clean_vader;
2) drugsComTrain_raw_vader;
3) drugsComTrain_clean_Textblob;
4) drugsComTrain_raw_Textblob.

Three new columns were added to each 4 dataset depending on the tool used for SA — TextBlob or VADER: (1) TextBlob_score or VADER_score: Polarity score from TextBlob or VADER, ranging from -1 (most negative) to 1 (most positive); (2) TextBlob_rescaled or VADER_rescaled: Rescaled TextBlob or VADER score ranging from 0 to 10; (3) TextBlob_label or VADER_label: Sentiment label for TextBlob or VADER scores (Negative, Neutral, Positive) based on the rescaled score.

Through this SA, we observed that VADER provided more distinct and nuanced results compared to TextBlob, Fig. 11. Specifically, TextBlob tended to classify a substantial portion of ratings as neutral, which could obscure insights in the data, Fig. 12. VADER's analysis, on the other hand, offered a better differentiation of sentiment, allowing for a more detailed understanding of user perceptions and experiences documented in the reviews, Fig. 13. Importantly, VADER distinguishes itself at identifying negative reviews, a crucial feature for the healthcare sector where prompt identification of adverse feedback is essential. This capability makes VADER a preferable tool in the contexts of this dataset, where distinguishing between subtle sentiment tones is crucial for data interpretation and decision-making. Both tools gave ratings less positive comparing with the users original rating. Since these ratings are based on SA and many drugs with high rating have reviews saying they worked but had side effects, it is possible to understand the discrepancy in values, Fig. 14.

V. Conclusion and Future Work

This study has systematically explored the UCI ML Drug Review dataset to explore the substantial roles that data visualisation and SA play in understanding patient feedback within the healthcare sector. By using tools like TextBlob and VADER, we have classified and analysed the sentiments expressed in patient reviews, which has allowed us to gain deeper insights into patient satisfaction and medication efficacy. Our findings underscore the potential of such analyses to enhance patient understanding and support more informed healthcare decisions. Moving forward, several avenues could

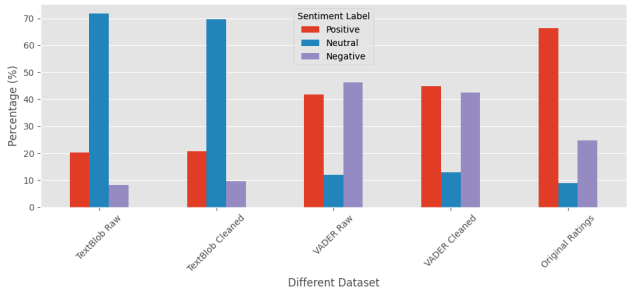

Fig. 11: Comparison of Sentiment Classifications and Ratings

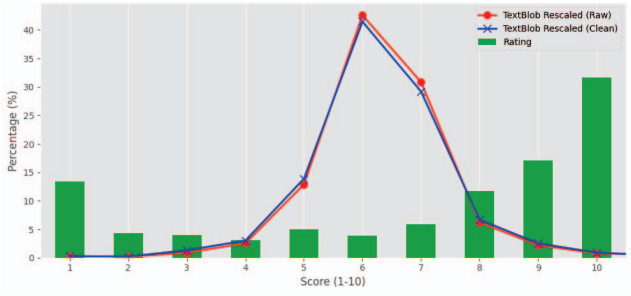

Fig. 12: Comparison of Ratings Frequency and TextBlob Rescaled Scores

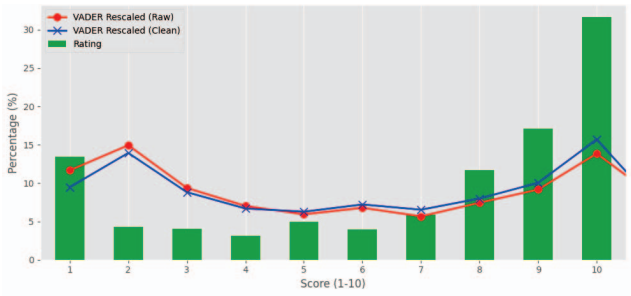

Fig. 13: Comparison of Ratings Frequency and VADER Rescaled Scores

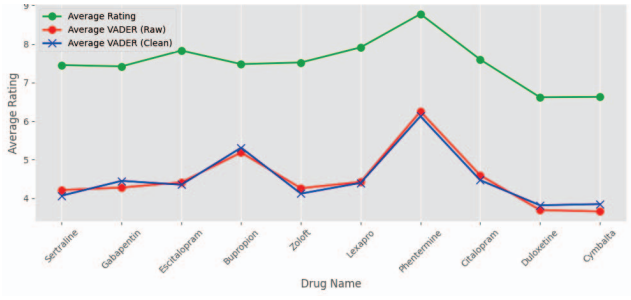

Fig. 14: Comparative Analysis of the Top 10 Drugs: Average Ratings vs. VADER Sentiment Scores

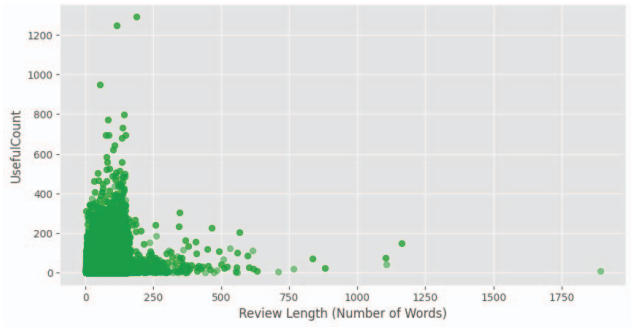

Fig. 9: Usefulness vs. Review Length

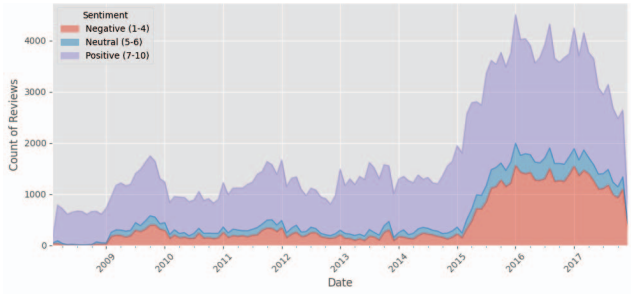

Fig. 10: Review Semtiment Over Time

further enrich our analysis and increase its applicability. Enhancements to the data pre-processing phase could refine the dataset, potentially revealing more nuanced insights. Additionally, exploring alternative SA tools may offer different perspectives on the data, broadening our understanding of patient sentiments. Finally, considering the integration of these datasets to support the development of a more targeted and effective drug recommendation system.

REFERENCES

[1] S. Kallumadi and F. Grer, "Drug Reviews (Drugs.com)." UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C5SK5S.

[2] S. Garg, "Drug recommendation system based on sentiment analysis of drug reviews using machine learning," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 175–181, IEEE, 2021.

[3] S. K. Mukhiya and U. Ahmed, *Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data*. Packt Publishing Ltd, 2020.

[4] B. Liu, "Sentiment analysis: A fascinating problem," in *Sentiment Analysis and Opinion Mining*, pp. 1–8, Springer, 2012.

[5] B. Panda, C. R. Panigrahi, and B. Pati, "Exploratory data analysis and sentiment analysis of drug reviews," *Computación y Sistemas*, vol. 26, no. 3, pp. 1191–1199, 2022.

[6] A. Chhabra, A. Sharma, K. Chhabra, and A. Chhabra, "A statistical analysis of sentiment over different social platforms on drug usage across high, middle and low-income countries," *Scalable Computing: Practice and Experience*, vol. 24, no. 4, pp. 971–984, 2023.

[7] Y. Qi and Z. Shabrina, "Sentiment analysis using twitter data: a comparative application of lexicon-and machine-learning-based approach," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 31, 2023.

[8] E. Lee, F. Rustam, H. F. Shahzad, P. B. Washington, A. Ishaq, and I. Ashraf, "Drug usage safety from drug reviews with hybrid machine learning approach.," *Computer Systems Science & Engineering*, vol. 46, no. 1, 2023.

[9] U. Ependi, A. Muzakir, and A. Wibowo, "Sentiment analysis on smart city mobile platform based on lexicon," in *2023 1st IEEE International Conference on Smart Technology (ICE-SMARTec)*, pp. 190–195, IEEE, 2023.