Drug Review Analysis Using Elasticsearch and Kibana

Monika Mishra

Department of Information Systems, California State University

Los Angeles

e-mail : [mmishra2@caltstatela.edu](mailto:mmishra2@caltstatela.edu)

**Abstract:** This paper aims to find useful insights from the drug review dataset. Reviews and ratings are an integral part of business in our digital age of consumerism. Through, this dataset, the most prevalent medical condition was found to be Birth Control. Popular and unpopular drugs, review count trend etc. will be found too . The machine learning technique will be used to detect anomalies in the average rating. A forecast of the average rating will also be made.

**1. Introduction**

Web-based reviews can be viewed as an orthogonal source of information for consumers, physicians, and drug manufacturers to assess the performance of a drug. This information can be leveraged to obtain valuable insights using data mining approaches. In this work we examine online user review data within the pharmaceutical field. Online user reviews in the pharmaceutical domain contain information related to multiple aspects such as effectiveness of drugs and side effects. Finding anomalies and predicting average rating is equally important to take corrective measures if required. In our paper, we have done both descriptive and predictive analysis of the drug review dataset using Elasticsearch and Kibana.

The dataset used for this project is from UCI ML repository. Drug Review dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating system reflecting overall patient satisfaction during the 2008-2017 period. The data was obtained by crawling online pharmaceutical review sites. The dataset is of size 84 MB with a TSV file format. It has 7 columns - id, drugName, condition, review, rating, date and usefulCount.

2. Related Work

Uysal [1] did a comparative performance analysis of techniques for automatic drug review classification. He used the classification machine learning technique to classify the drug review as positive and negative based on different fields using Support Vector Machine and Naïve Bayes algorithm. . My work is different as I have performed both descriptive and predictive analysis. And in predictive analysis, I have emphasized on finding anomalies in average rating and forecasting the average rating.

Bhargava’s [2] research , too, was based on the drug review but his approach is different than mine. While my work is based on descriptive analysis and forecasting of average rate, his work is to use unsupervised clustering model to cluster drugs based on similar usage and benefit. He used k-means clustering algorithm, agglomerative clustering and BIRCH (balanced iterative reducing and clustering using hierarchies) model to cluster drugs and analyzed the model using quantitative performance metrics (Silhouette Score and Calinski-Harabasz Score), cluster sizes, as well as visualization by the means of scatter plot of clusters, distances from final centroids, word clouds and dendrogram.

3. Architectural Workflow

The dataset was downloaded from UCI ML repository. In the data pre-processing stage, the *date* column was changed from date to timestamp format as machine learning on Kibana can only be done on the dataset containing timestamp data. Mapping changes were done during data uploading process. The *type* of *drugName* and *review* field were changed from *text* to *keyword*. Visualizations were created using *Visualize* and *Canvas* feature in Kibana. At last, anomaly detection and forecasting of average rating were done using the *Machine* *Learning* feature.

**A close up of a sign

Description automatically generated**

Fig 1. Architectural workflow

3. Platform Specifications

The platform specification for the project is provided in the **Table 1**.

|  |  |
| --- | --- |
| **Elasticsearch & Kibana** | **OS – Mac OS X** |
| Version – 7.2.0  No. of Node – 1  Disc Size – 456.6 GB  Memory Size – 1.4 GB | Processor Speed – 3.1 GHz  Memory – 8GB  No. of Processor – 1  No. of cores - 2 |

Table 1. Platform specification

4. Background/Existing Work

The project work is based on the lab works done in the class. All the visualizations are inspired from the lab work. In one of the labs, a pie chart was created for the bank data after it was uploaded to Kibana using curl command. The size of the slice was based on the record count while the bucket was split based on the balance and aggregated as range. A sub bucket was added too based on the age. Similarly, a mapping for Shakespeare data was created and then the data was uploaded to Elasticsearch. A bar chart was created for the unique count of the play name.

The Machine Learning part of the project is based on the example provided by the Elasticsearch team [3] where they created a single metric job to find anomaly in the average taxi trip duration on the New York taxi trip data. They also performed the forecast of average taxi trip duration.

5. Our Work

We implemented various visualization using Kibana *Visualize* and *Canvas* feature to derive important insights from the dataset.

A close up of a logo

Description automatically generated

Fig 2. Rating distribution

**Fig 2.** shows the overall rating distribution of the drugs on pharmaceutical site. Most of the drugs have been rated 10 (31.61%) and 9 (17.07%).

A close up of a map

Description automatically generated

Fig 3. Yearly review count trend

**Fig 3.** shows the yearly review count. Over the years the review count has seen both an increase and decrease. But from 2015, it has shown a dramatic increase in the review count.

A screenshot of a cell phone

Description automatically generated

Fig 4. Most common medical conditions

**Fig 4.** shows the most common medical conditions among the patients. Birth Control, Depression, Anxiety, Pain are some of the most common conditions.

A screenshot of a cell phone

Description automatically generated

Fig 5. Number of drugs per condition

According to **Fig 5.**, Pain, Birth Control, High Blood Pressure are some of the medical conditions which have the highest number of curable medicines.

A close up of a logo

Description automatically generated

Fig 6. Most reviewed drugs

**Fig 6.** displays the most reviewed drugs in the dataset. Levonorgestrel drug has been reviewed the most.

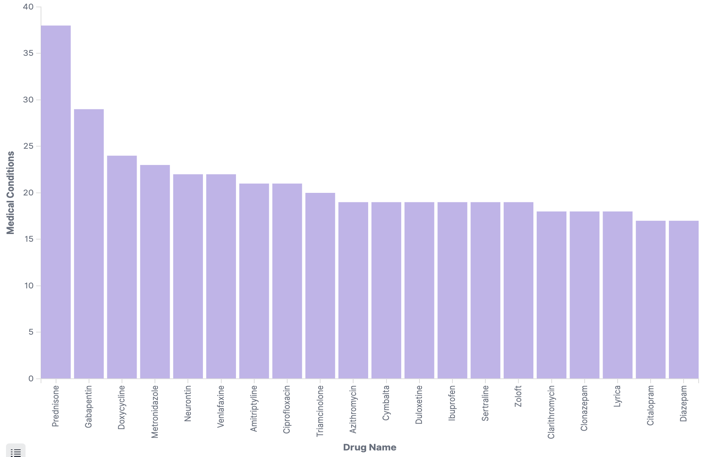


Fig 7. Drugs used in multiple medical conditions

There are certain drugs that are used in multiple conditions. **Fig 7.** shows the same. From further analysis it was found that a medicine can be used for diverse treatment. For example, Prednisone drug is used to cure multiple medical conditions.

A screenshot of a cell phone

Description automatically generated

Fig 8. Conditions cured by the drug - Prednisone

Prednisone drug is used to cure a variety of medical conditions such as inflammatory conditions, asthma, different types of arthritis, skin diseases such as psoriasis, skin rash, leukemia etc. This is shown in **Fig 8.**

A screenshot of a cell phone

Description automatically generated

Fig 9. Significant drugs with the best rating

**Fig 9.** shows the top significant drugs with the best average rating. By the term significant means, we have considered only those drugs whose review count is greater than 30. This visualization has been created by using the *Canvas* feature of Kibana. The below SQL was used to get the desired result:

**select** **avg**(rating) "rating" ,

drugName,

**count**(drugName) "cnt"

**from** drugs

**group by** drugName

**having** cnt >30

**order by** rating **desc**

**limit** 20

A screenshot of a cell phone

Description automatically generated

Fig 10. Significant drugs with the worst rating

**Fig 10.** shows the significant drugs with the worst average rating. By the term significant means, we have considered only those drugs whose review count is greater than 30. This visualization, too, has been created by using the *Canvas* feature. The below SQL was used to get the desired result:

**select** **avg**(rating) "rating" ,

drugName,

**count**(drugName) "cnt"

**from** drugs

**group by** drugName

**having** cnt >30

**order by** rating **asc**

**limit** 20

A screenshot of a cell phone

Description automatically generated

Fig 11. Highest Reviewed Medical Conditions

**Fig 11.** shows the five most reviewed medical conditions each year. It can be seen that depression and anxiety are the consistent medical conditions present in every year.

It is also worth noting that the birth control condition saw a dramatic spike in the year 2015.

A screenshot of a computer

Description automatically generated

Fig 12. Anomaly detection of average rating

The*Machine Learning* feature of the Kibana is used to detect the anomaly. **Fig 12.** shows the anomaly detection of the average rating over the years. There was just 2 critical anomalies and 2 major anomalies.

**A screenshot of a computer

Description automatically generated**

Fig 13. Forecasting of average rating

The forecast of the average rating can also be made.The yellow line in the **Fig 13.** shows the same. The yellow line shows the upper and lower bound of the average rating.

**6. Conclusion**

Following conclusion can be drawn from the project:

* Birth Control followed by Depression , Anxiety, and Pain are the top problems faced by the people.
* Maximum drugs have been rated 10.
* People’s reactions are extreme.
* One drug can be used to cure multiple medical conditions.
* Depression and Anxiety medical conditions comes under top 5 medical conditions in all the years being analyzed.
* Review count has dramatically increased since 2015.
* Medical conditions “Pain” has the highest number of curable medicines available.
* Levonogestrel drug is the most reviewed drug.
* Birth Control medical condition is pretty dominant in all the years being analyzed.
* Since 2015, Birth Control condition is spiked up.
* Elasticsearch and Kibana can be used for anomaly detection and forecasting.

### References

[1] Uysal, A. K. Comparative Performance Analysis of Techniques for Automatic Drug Review Classification. *Celal Bayar Üniversitesi Fen Bilimleri Dergisi*, *14*(4), 485-490.

[2] Bhargava, A. (2019). Grouping of Medicinal Drugs Used for Similar Symptoms by Mining Clusters from Drug Benefits Reviews. *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur - India, February 26-28, 2019.*

[3] https://www.youtube.com/watch?v=wJVgh5knV4E

[4] Github Link: https://github.com/monika2403/mmishra2/tree/master/CIS%205560

[5] Dataset Link: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

[6] Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018, April). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 International Conference on Digital Health* (pp. 121-125). ACM.