A Real Time Drug Analysis System Using Efficient Machine Learning Algorithms

Likitha A B[1], Venkatesh U C[2], Naseema C A[3], Divya A K[4]
[1, 2, 3, 4] Dept of Computer Science and Engineering
[1, 2, 3, 4] KVG College Of Engineering,Sullia

Abstract- *Virtual health communities provide a wealth of medical information valuable to medical practitioners, system admins, patients .Our system collects real-time medical reports from reliable sources, where patients share their perspectives and side effects of the drugs they use. We aim to summarize user posts for each drug, offering concise conclusions for both the medical practitioners and patients. Additionally, we plan to classify users based on their emotional states and conduct knowledge exploration through association learning to uncover patterns relevant to the triad of drugs, symptoms, and treatments.*

Keywords- Data Science, Symptoms, Diseases, Drugs, Machine Learning, Eclat, Lesk Based, Algorithm

I. INTRODUCTION

Alongside the exponential growth of the web, the amount of electronic data is increasing significantly. While this abundance is beneficial in the Information Age, it also poses challenges in relation to terms of time, storage, and understandability. Extracting meaningful knowledge from this vast information remains a significant challenge.

To address this, we propose applying key operations such as Association analysis, Summarization, and Sentiment Analysis in data sourced from the health forum site health-boards.com extracting insights from of health posts.

Summarization involves extracting and condensing information from a source to present the premier valuable providing content to the user in a form that meets their specific application needs [1].

Summarization plays a vital role in various NLP applications, including Information Retrieval, Quality Analysis, and Text Comprehension. There are two primary types of summaries: Extractive, which reuses words and sentences directly from the text, and Abstractive, which involves the regeneration of the extracted content [2].

Association rule mining is a popular and widely recognized machine learning task used to uncover interesting relationships between variables in large databases. The rules generated through association mining consist of two disjoint sets of items, formatted as LHS (Left Hand Side) => RHS (Right Hand Side). These rules indicate that the Whenever the LHS set occurs, RHS is likely to follow. The discovering association patterns involves two steps:

1. Association Rule Generation
2. Interesting Rule Selection

After generating the rules, they are extracted and undergo post-processing. The extracted rules from the health boards dataset may manifest in one or more of the following forms:

1. symptoms->disease
2. disease->disease
3. medicine->disease
4. disease->medicines
5. Age group->disease.

Sentiment Analysis (SA) involves extracting sentiments from text, which can encompass opinions, attitudes, Sentiments regarding an entity. This entity could be individuals, events, or topics, often covered in reviews. Walaa Medhat regarded Sentiment Analysis as a classification process with three levels: Document level, Sentence level, and Aspect level. In performing SA, important features are first selected from the text, and then classification is carried out using an appropriate classifier.

In our case, we are examining feedback from health posts where the subject entity isa drug. Therefore, our categorization is under the aspect level.

II. RELATED WORKS

Survey Papers

1. IEEE PAPER TITLE: Association Rule Mining on Type 2 Diabetes using FP-growth association rule
YEAR OF PUBLICATION: 2013
AUTHORS: Nandita Rane1, Madhuri Rao

METHODOLOGY: FP growth algorithm used
LIMITATIONS: this paper mainly concentrates on only diabetes disease and related patterns.

2. IEEE PAPER TITLE: A Comparative approach for Sentiment Analysis from Summarized User Health posts
YEAR OF PUBLICATION: 2016
AUTHORS: Mr Ajay A V, Dr.Chethan H K
METHODOLOGY: NLP techniques used
LIMITATIONS: this paper aims at comparative analysis of different sentiment analysis techniques which is employed to process the patient's opinions. Doesn't predict the connection between symptoms-diseases-drugs.

3. IEEE PAPER TITLE: Knowledge discovery from user health posts
YEAR OF PUBLICATION: 2015
AUTHORS: Vinod L. Mane, Suja S. Panicker, Vidya B. Patil
Methodology USED: AIT algorithm, FP growth algorithms and apriori used
LIMITATIONS: The used algorithm takes more time for processing, less efficient.

Gap Analysis

- Identification of relationship b/w symptoms, diseases and drugs.
- More efficient algorithm used.
- ECLAT algorithm used.
- Summarization (Lesk based algorithm used)
- Pattern prediction (Eclat Algorithm used)
- Sentimental analysis (Lesk based algorithm used)

III. PROPOSED WORK

The proposed system collects real-real-time health updates from reputable websites, wherein patients share their encounters and their experiences and side effects of the drugs they have utilized. It summarizes user posts for each drug, providing beneficial insights for both the medical field and patients at first glance. Additionally, the system performs knowledge extraction from user posts, identifying valuable patterns in the triad of drugs, symptoms, and treatments using Association Rule Mining.

3.1 Objectives

- To summarize all user posts, where patients articulate their perspectives, experiences, and side effects of the drugs they have used, and to derive useful conclusions from them.
- Proposed system is a medical sector application.
- Proposed system collects the posts from the users (medical practitioners, patients) related to side effects on drug.
- The proposed system summarizes all user posts and derives useful conclusions from them.
- Proposed system discovers useful patterns based on side effects per drug.
- The proposed system utilizes "Association Rules" to uncover patterns.
- The proposed system is a novel online community that provides a vast array of medical information beneficial for healthcare professionals, patients, and others.

3.2 Advantages

- This work will assist patients in discovering the associations among various drugs, diseases, and symptoms.
- It will assist doctors in identifying side effects of various drugs, enabling them to prescribe more effectively for other patients with similar conditions.
- Pharmaceutical companies will also benefit from our system, as it classifies users of a particular drug into categories such as depressed and satisfied.
- This will indirectly inform companies about which drugs are popular and whether there is a need to produce alternative medications.
- Thus, our work will equally benefit all three parties: the medical fraternity, the patient community, and pharmaceutical companies.

IV. METHODOLOGY

4.1 Unsupervised Learning

Step 1: In the 1st step we collect the medical data (patient opinions [feedback] based on drugs). We referred sources like www.patientslikeme.com,www.healthboard.com and www.kaggle.com.

Step 2: Patient opinions are then preprocessed and irrelevant data removed and only relevant data extracted and inputted to the algorithms.

Step 3: Then we input the necessary things required for algorithms, initially preprocessed patients opinions are summarized using lesk based algorithm and the outcome of lesk based algorithm is inputted to Éclat algorithm to uncover to discover the medical patterns (symptoms-diseases-drugs).
Step 4: The Éclat algorithm will uncover medical patterns that reveal relationships between symptoms and symptoms,

symptoms and diseases, diseases and diseases, and diseases and drugs.

Step 5: Medical patterns displayed on GUI (front end).

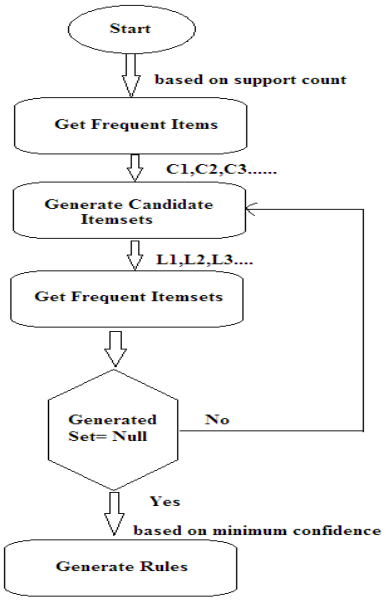
Step 6: The results of the data science algorithms are analyzed and visually represented.

4.2Eclat Algorithm

4.2.1 Steps

1. *count(s) of each item. Add the transaction ids instead of specifying the actual support.*
2. *Generate L1 by comparing with minimum support count.*
3. *Use Lk-1 and join it with itself to create the set of potential k-item sets.*
4. *Review the candidate k-item sets and compute the support for each candidate k-item set. When we find support count of candidate items, we compare with previous step, no need to again query the database and compare with original data-set.*
5. *Add to the frequent item set until C becomes a null set.*
6. *For each item in the frequent item set, create all non-empty subsets.*
7. *For each non-empty subset, calculate its confidence. If the confidence meets or exceeds the specified threshold, include it in the Strong Association Rule set.*

4.2.2 Eclat Algorithm Flow



Flow of the Algorithm

Fig: 3 – Eclat Algorithm Flow Diagram

4.2.3 Eclat Algorithm Pseudo-code

```
double.Parse(lv_Transactions.Items.Count.ToString()) / 20;
double _supportCnt = (double.Parse(lv_Transactions.Items.Count.ToString()) / 100) * 10;
double _confidence = 0.9;
        //function to calculate the L1 and store in buffer
        Dictionary<string, string> _FrequentItemsMain = L1(_supportCnt);
        Dictionary<string, string> _Candidates = new Dictionary<string, string>();
do
        {
          //function to calculate C2,C3.... and store in buffer
          _Candidates = GenerateCandidates(_FrequentItemsMain);
          //function to calculate L2,L3.... and store in buffer
          _FrequentItemsMain = GetFrequentItems(_Candidates, _supportCnt);
        }
while (_Candidates.Count != 0);
        //function to generate rules or patterns
        List<ClassRules>RulesList = GenerateRules();
        //compare with the confidence and find strong rules
        List<ClassRules>StrongRules = GetStrongRules(_confidence, RulesList);
        //function to display the final results (L and Patterns)
        Result(DictionaryAllFrequentItems, StrongRules);
```

4.2.4 Lesk based Algorithm

Algorithm Steps

- Retrieve data from the Opinion Database.
- Scan word-net (collection of all symptoms, diseases, Drugs and SA Words)
- for each entry Ui[Opinions] in buffer[storage server] do
- Trace all Keywords, using the following steps
    - Tokenization [keyword extraction method – removing the stop words and retrieving the keywords]
    - Remove punctuation, special characters, number etc..
    - Clustering the Keywords (grouping of similar objects)
        - by comparing with the predefined data-set (created by the admin)
        - String Comparison and Identify the Symptoms, Diseases, Drugs and Positive and Negative Words.

✧ Output - Summarized Results

V. EXPERIMENT RESULTS

ECLAT ALGORITHM

Performance Factor

Data Structure – array based

- Memory Utilization – It varies based upon data set [less for small datasets]

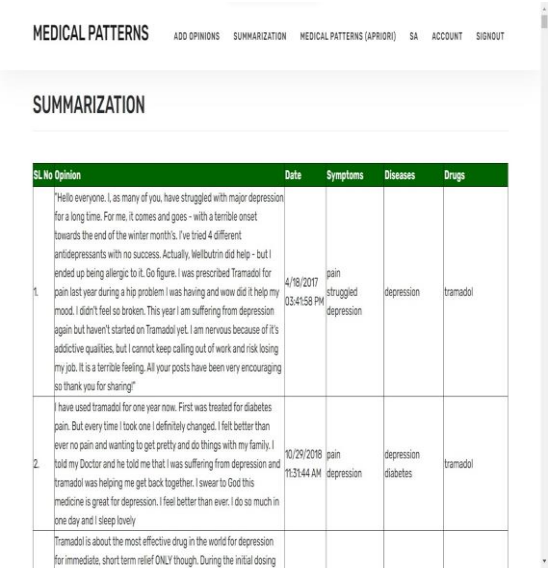
No.of.scans - single scan required

Execution time - execution time depends on producing candidates

No of Instances (records)	Execution Time (milliSecs)
Around 2k	1557
Around 1k	1295
Around 500 records	865
100 records	445

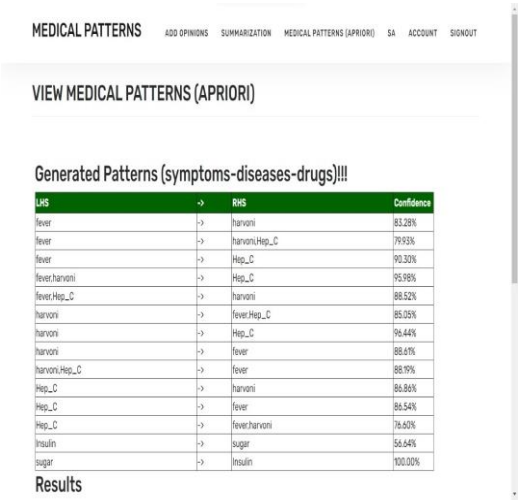
Summarization Module (View)

This core module summarizes all user posts to derive useful information or conclusions. It utilizes a keyword extraction method to extract symptoms, diseases, and drugs from user health .



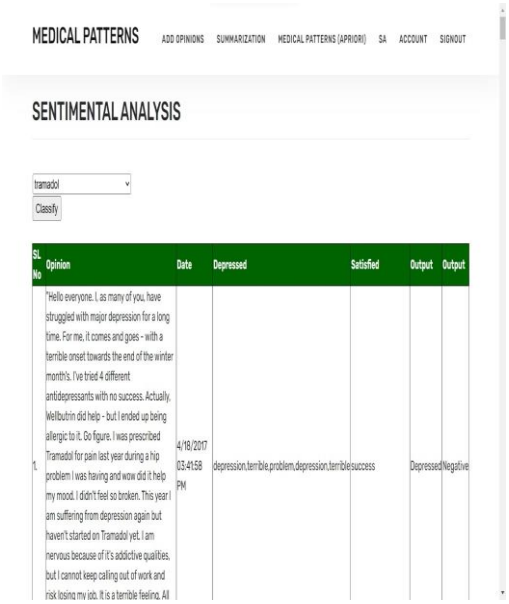
Pattern Prediction Module (View)

This core module of the project discovers meaningful patterns based on user-posted side-effects per drug. It identifies correlations between symptoms, diseases, and drugs. To predict these relationships, we employ the efficient "Eclat Algorithm" for pattern discovery.



SA Module

In this module, patients provide ratings for specific drugs, allowing us to track and analyze the top-rated drugs and their details.



VI. CONCLUSION

In this project, we gather current health articles from credible websites and use data mining techniques to uncover various associations among diseases, drugs, and symptoms. Through Association Rules algorithms, we conduct knowledge extraction from user posts to detect valuable patterns. This approach assists doctors in identifying drug side-effects and improves their ability to prescribe more effective treatments

for patients with similar conditions. Pharmaceutical companies benefit by gaining insights into drug responses and popularity, aiding decisions on production. Patients benefit by accessing user opinions, enhancing their ability to make informed decisions about medications and increasing awareness of drug side-effects experienced by others.

Future Enhancements

- We can use more algorithms for medical patterns prediction and can compare the algorithm results and can identify the better algorithm.
- We can add more drugs.
- We can add more disease types.

REFERENCES

[1] JayashreeR,Srikanta Murthy K,Basavaraj .S.Anami, "Categorized Text Document Summarization in the Kannada Language by Sentence Ranking", 12th International Conference on Intelligent Systems Design and Applications (ISDA), pp 776-781, 2012.
[2] AlokRanjan Pal, DigantaSaha, "An Approach to Automatic Text Summarization using WordNet", IEEE International Advance Computing Conference (IACC), 2014.
[3] JesminNahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen, "Association rule mining to detect factors which contribute to heart disease in males and females", J. Nahar et al. / Expert Systems with Applications 40 (2013) 1086–1093, Elsevier, 2012.
[4] Lakshmi K.S, G. Santhosh Kumar, "Association Rule Extraction from Medical Transcripts of Diabetic Patients",IEEE,2014.
[5] WalaaMedhat, Ahmed Hassan, HodaKorashy, "Sentiment analysis algorithms and applications: A survey", In press, Elsevier, 2014.
[6] Rafael Ferreira, FredericoFreitas, Luciano de Souza Cabral, Rafael DueireLins, Rinaldo Lima, Gabriel Franca, Steven J. Simske, and Luciano Favaro, "A Context Based Text Summarization System",11th IAPR International Workshop on Document Analysis Systems,pp 66-70, 2014.
[7] C. Lakshmi Devasenal and M. Hemalatha, "Automatic Text Categorization and Summarization using Rule Reduction", IEEE- International Conference On Advances In Engineering, Science And Management (ICAESM -2012), pp 594-598, 2012.
[8] Sara Keretna, CheePeng Lim, Doug Creighton, "A Hybrid Model for Named Entity Recognition Using Unstructured Medical Text", Proc.Of the 2014 9th International Conference on System ofSystems Engineering (SOSE), Adelaide, Australia- June 9-13, pp 85-90, 2014.
[9] SaeedMohajeri,AfsanehEsteki, Osmar R. Zaiane and DavoodRafiei, "Innovative Navigation of Health Discussion Forums based on Relationship Extraction and Medical Ontologies",IEEE International Conference on Bioinformatics and Biomedicine, pp 13-14, 2013.