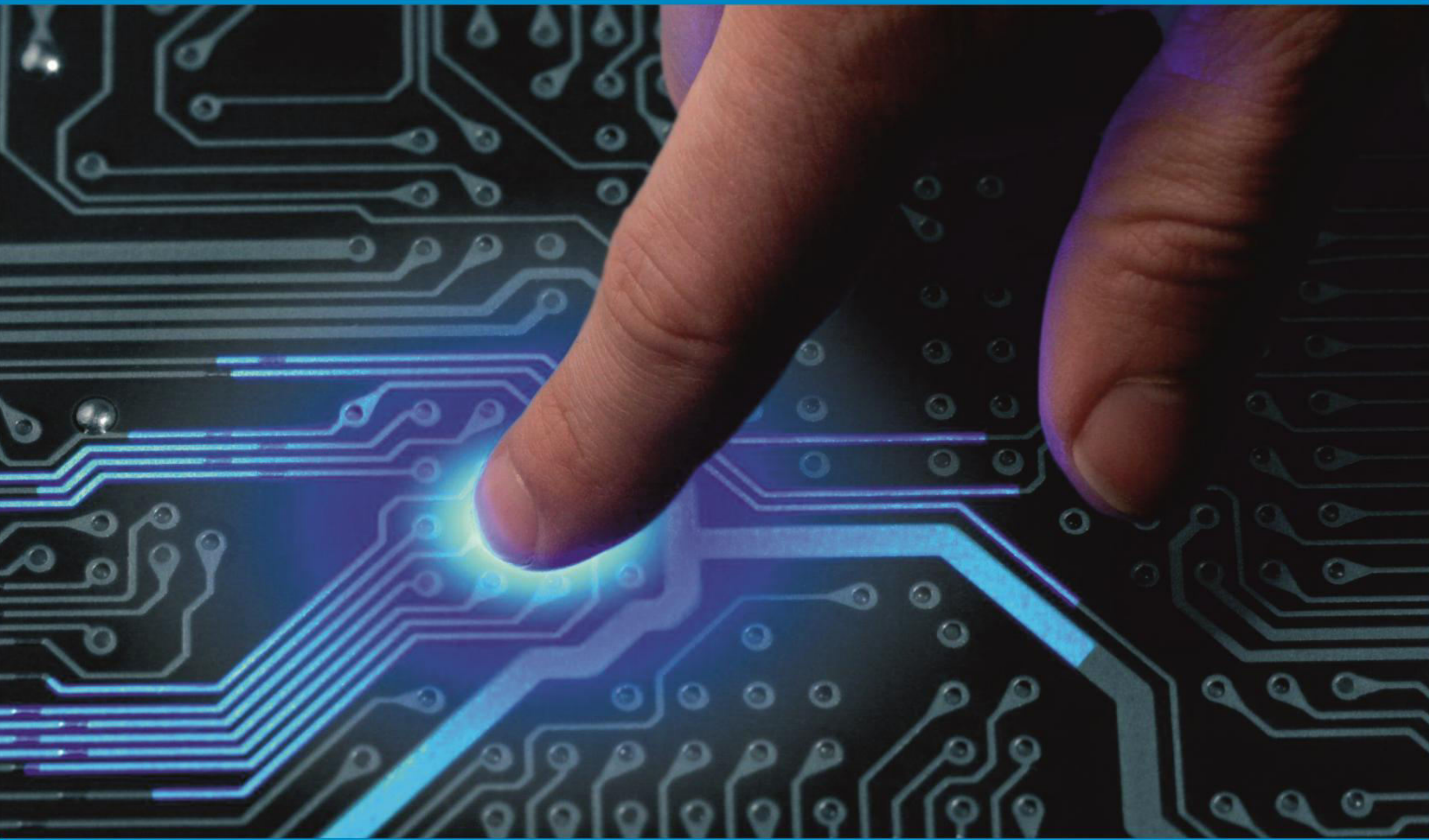
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 4, April 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

IJIRCCE

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904032 |

Drug Analysis Using Sentiment Analysis

Preetesh Purohit[1], Khushbu Goyal[2], Rohit Sharma[3], Vaishnavi Kadambari[4], Vani Saraf[5]

Department of Computer Science & Engineering, Medi-Caps University, Madhya Pradesh, India [1, 2, 3, 4, 5]

ABSTRACT: The advancement of medicine introduced a health and environmental transformation. Such drugs are useful only if they are free from contamination and are well handled. Different chemicals and experimental approaches for the evaluation of drugs have been established at frequent levels to determine medicinal products' suitability. These drugs can produce contaminants at different phases of growth, storage and distribution that makes the medication dangerous to prescribe and therefore must be identified and quantified. A significant part is played by analysis modulation and procedures. This analysis emphasises the significance of the analytical techniques and the testing approaches in evaluating the consistency of the drugs.

There are various internet reviews of medicines from patients. This analysis offers a short description of approaches to drug exploration factor mining. In small pre-marketing research studies, several adverse drug responses to chronic conditions were not discovered. It was noticed only after a lengthy post-marketing drug use survey. A significant area of study for the pharma companies is the prediction of drug reactions as quickly as possible. The key obstacle is to extract serious principles from brief and disruptive comments.

KEYWORDS: Drug Analysis, Medicines,Sentiment Analysis, Classification, EDA, Confusion Matrix, nltk, Word Cloud.

I. INTRODUCTION

Sentiment analysis is a very important activity of natural language processing, with the goal to determining people's feelings, beliefs, and behaviours toward goods, programmes, people, organisations, and other institutions [13]. Aspect-level sentiment analysis is a fine-grained sentiment analysis task that explores the polarity of a particular aspect in a sentence [2]. For instance, in the sentence "*This medication works well for water retention, but its side effect is serious*", the sentiment polarity for the aspects "water retention" and "its side effect" is positive and negative, respectively. The key explanation for domain dependency in aspect-level sentiment classification is that a term can have several meanings. If a goal is identified, the pharmaceutical companies and, quite significantly, several research centres will have a simplified series of early processes to find molecules with the right characteristics to produce good drugs. The key goal of the research is the establishment of general techniques and concerns in analytical techniques to isolate, classify and measure active pharmaceutical ingredients (APIs) which can be used in different roles in the drug spectrum. The analysis also examines the problems and criteria which need to be investigated as empirical approaches are validated. [13]

The mechanism has evolved to the point that medications are crafted instead of found. To reveal, improve and generate pharmaceutical products, the production and verification of analytical techniques is of significant importance.

II. RELATED WORK

Previous findings have demonstrated that viewer content from a well-being viewpoint is valuable. [8] Jane Sarasohn-Kahn wrote one of the standard paper in this domain. It says people also search for "patients like them" reports on the web that is difficult to locate among families and friends. For the last 10 years, researchers have analysed the emotional effect of user interface and the seriousness of negative drug interactions by collecting sentimental and textual knowledge.

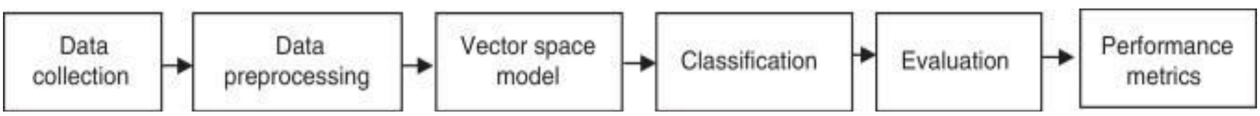


Fig.1: Problem design [8]

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904032 |

III. DATASET STATISTICS

This Dataset includes patient feedback on individual medications, associated disorders, as well as a 10-star patient satisfaction index. Crawling online pharmaceutical review pages has yielded this information. These features include 'drug name', which is the drug's name, 'condition', which is the patient's condition,'review', which is the patient's assessment, 'rating', which is the patient's 10-star rating for the medication, 'date', which is the date of the entry, and 'useful count', which is the number of consumers that found the review helpful. The focus variable here is the review's sentiment, which must be the same as expected. The drug name and condition are categorical features, while the date is a date property. The ranking and user count are quantitative characteristics, and the analysis is a qualitative feature.[3][4][5]

a. Analyze exploratory data

The aim of exploratory data analysis (EDA) is to gain insight into the data and summarise the most significant characteristics. To comprehend the functions' interdependence or association.

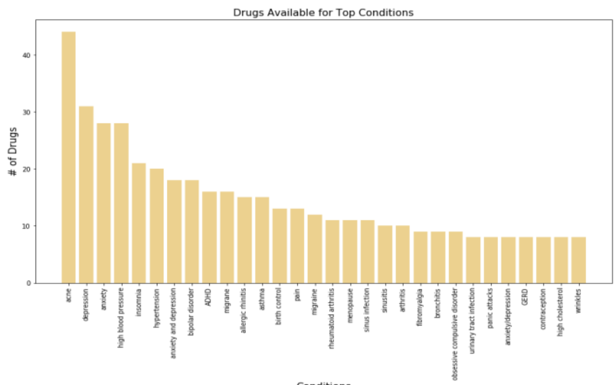


Fig.2: Number of conditions present per drug

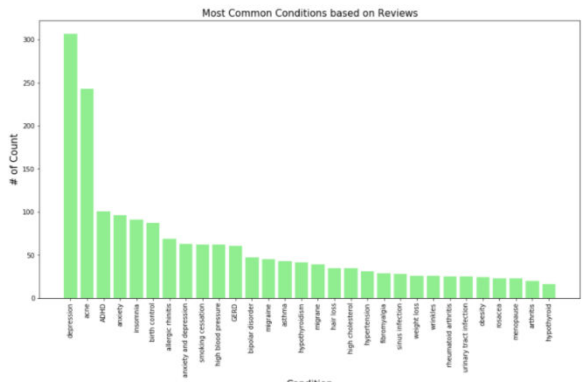


Fig. 3: Most Common Conditions based on Review

V. SENTIMENT ANALYSIS ON USER REVIEWS

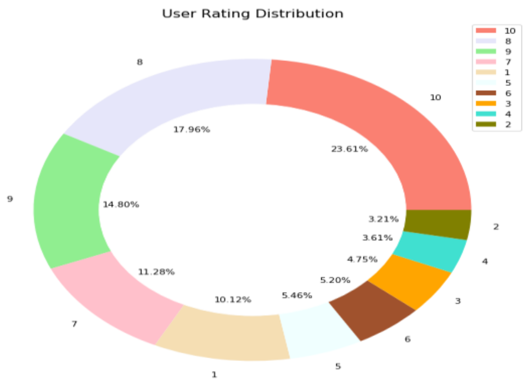


Fig. 4: User Rating Distribution

```
1895    nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
1896    nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
1897    nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
1898    nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
1899    nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
                                                       ...
120616  nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
120617  nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
120618  nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
120619  nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
120620  nausea , diarrhea , dizziness , lightheadedness , headache , weakness , or trouble sleeping may occur.
Name: Side_Effects, Length: 312, dtype: object
```

Fig. 5: Side Effects

Word Cloud: "Keyword enrichment analysis" is done by the Word Cloud. It calculates keyword enrichment through all contrasts for a specific keyword[7].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904032 |

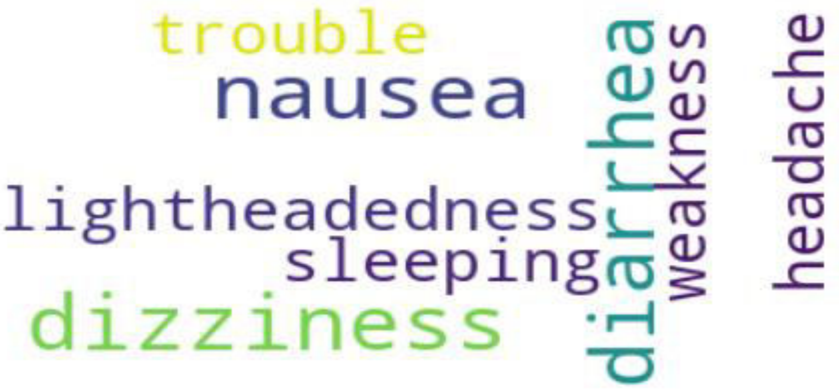


Fig. 6: Word Cloud Visualization of the dataset

VI. COLLECTION AND DATA SOURCES

The first step is to obtain data sets in the science data pipeline. We have separated 3 healthcare directories (Drugs.com, WebMD, DrugLib) containing information about medications such as prescribed medications, health problems, age, sexuality and other conditions. [3][4][5]

- We built a Scrapy web crawler that scrapes WebMD with Spider. Attributes such as gender were possible to isolate because textual characteristics varied within each person.
- DrugLib& Drugs.com were scrapped using BeautifulSoup and request — an HTML module that can be processed and downloaded. The first and only challenge was to remove age from the definition of the customer. That's why, we now have a phrase regularly to solve the problem.

VII. DATA CLEANING AND INTEGRATION

a. Filtering and cleaning: Even as we can observe through the table in typically the data collection segment, DrugLib has about 1500 conditions between 4200 reviews which often would seem sporadic in comparison with other info sources[3]. That is usually for the reason that scraped info was the case very sensitive, had a lot of misspelt conditions in addition to multiple conditions have been grouped into a single ("anxiety, panic episodes, nervousness"). To fix typically the above issues around all 3 info sources, we executed methods like switching to lowercase, cutting off in addition to filtered unnecessary circumstances with <span> plus "Not Listed" pointed out within them.[3],[4],[5]

b. Jaccard and Fuzzy Matching: Drug names and symptoms ranged widely across three information sources and therefore, must be transformed using Outcome Assessment methods. Jaccard technique of resemblance is used to search several sets (drug, condition).

c. Pre-processing and retrieval of data:
Machine learning models hardly function with original text input. So the text should be translated to graphical values or, more simply, to numerical variables. Our method to extract features that explain the existence of terms in a text is Bag-of-Words (BoW). Two separate ways to transform the text into a vector have been experimented with.

d. Model predictive:
We have tested several models and deep learning methods for reviewing the feeling of the text into three main categories (Positive, Neutral, and Negative).[6] By classifying the ranking in the following way, we established target labels for emotion classification:

Rating	Target Label
>=7	**Positive**
>= and,7	**Neutral**
<=4	**Negative**

Fig. 7: Model Predictive Labels

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904032 |

VIII. METHODOLOGY

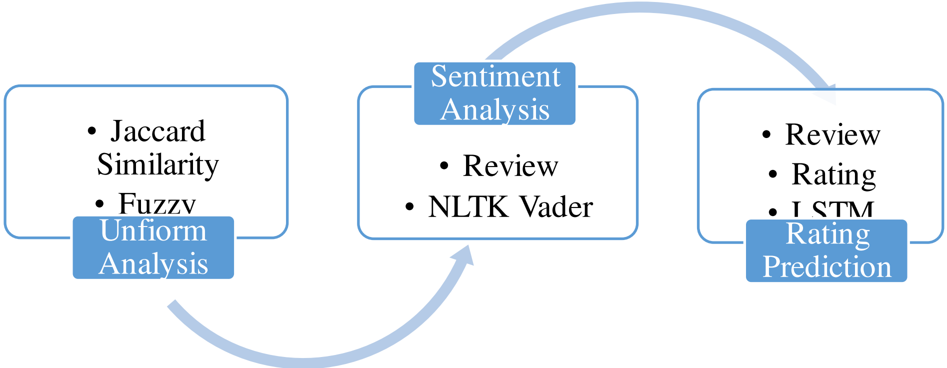


Fig.8: Process Explained from RAW Data Extraction to Drug Analysis

A. Sentiment Classification using LSTM model:

For analysing consumer views, emotion classification methods have been commonly used. Hand-crafted features are required in traditional supervised learning approaches, which necessitates a detailed understanding of the domain.[12] Since user feedbacks are typically brief, there aren't many mechanisms for successful categorization. Term embedding models may thus have been used to explore new word meanings in several contexts.[2]

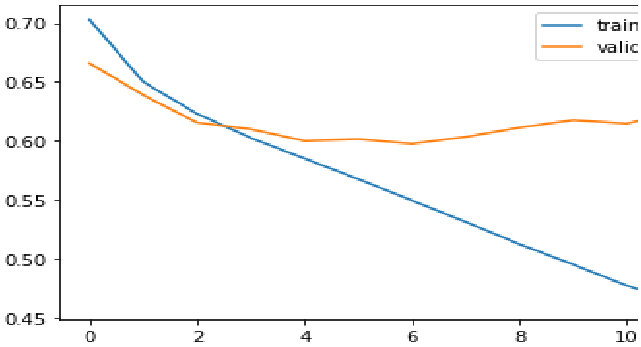


Fig. 9: Training & validation

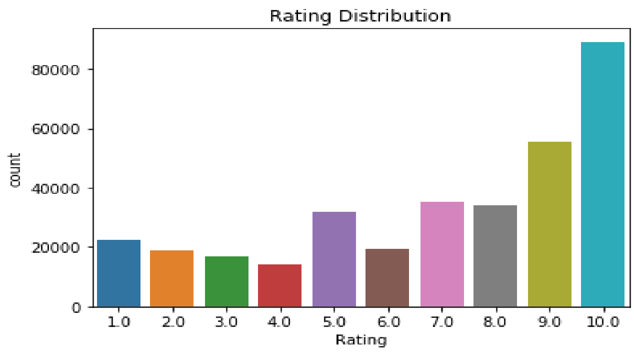


Fig. 10: Rating Distribution

B. EXAMINATION OF VADER SENTIMENTS:

We've used the sentiment analyser NLTK VADER (Valence Aware Dictionary and Sentiment Reasoning), a method that analyses the dictionary and rules. Since most of the drug comments are written in casual language, VADER is a crucial part of non-technical, social media and product reviews. Any term in the lexicon, then, is evaluated as positive, negative, or neutral. The score is a metric, which measures and normalises the sum of all lexicon scores between -1 and 1[8].

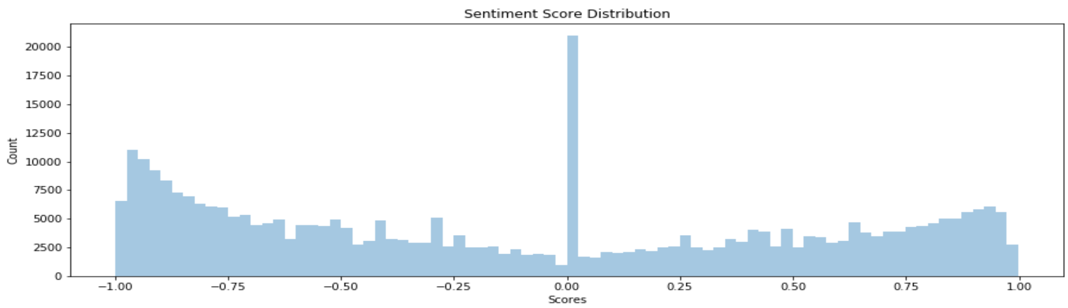


Fig. 11: Sentiment Score Distribution

C. CLASSIFICATION OF FEELINGS BY STATISTICAL MODELS:

We chose Model word sacks, instead of word embeddings, because it fits best where the datasets are small and the meaning is very domain-specific.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904032 |

- **TF-IDF:** We used NLTK's pre-trained word tokenizer, snowball stemmer and unigram and bigram templates in the text pre-processing. We have decided to use Snowball Stemmer (improved Porter Stemmer version) because it prevents overfishing, provides tempo, accuracy, and support for several languages.
- **Hashing:** In our situation, the TF-IDF works decently, but carries two limitations. Firstly, the language can become very broad and does not effectively recapture it. Secondly, converging takes longer. Thus, we tried the Hashing vectorizer to solve these shortcomings by using hacks as numerical indexes to encrypt tokens. Finally, the accuracy of the models using the approaches discussed above was compared in the assessment.[10]
- **ML simulations:** We used Randomised Search CV for tuning parameters to attain higher precision in tree-base models by drawing a random value from the set of values for each hyperparameter searched and testing the model for such hyperparameters for each iteration.

D. PREDICTIVE MODELS PERFORMANCE EVALUATION:
Long Short Term Memory - LSTM Model:

```
Y data
(269882, 10)
[[0 0 0 ... 0 1 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]
 [0 0 0 ... 0 0 1]]
```

Fig. 12: Y contains Rating's Values

```
X test data
[[   0    0    0 ...   18 1271 1572]
 [   0    0    0 ...  282 1496  133]
 [   0    0    0 ...   44    9    4]
 ...
 [   0    0    0 ...  234  109  130]
 [   0    0    0 ...  107  382   56]
 [   0    0    0 ...   57   10 1681]]
(67471, 200)
```

Fig.13: X contains Review's Values

E. Predictive Models Entire:

	Reviews	Rating	Website
0	"my son is halfway through his fourth week of ...	8	drugs.com
1	"the actavis generic version of this medicatio...	1	drugs.com
2	"my son was just diagnosed adhd today. he's 5 ...	7	drugs.com
3	"the first few days on 1 mg in the morning, he...	4	drugs.com
4	"ours 8 year old son as done so much better wi...	10	drugs.com
...	...	...	...
337332	i am having some of these side effects and wil...	2	webmd
337333	made me tired,achy,and was told not to take st...	2	webmd
337334	it makes me feel like crap after i take it.\ni...	3	webmd
337335	when i strated taking the medication i was fin...	2	webmd
337336	due to all the medication i'm was unable to co...	3	webmd

Fig.14: List of Review & Rating

	anger	anticipation	disgust	fear	joy	negative	positive	sadness
0	0	5	0	2	2	3	3	1
1	0	0	0	1	0	4	0	1
2	1	4	0	1	3	2	3	1
3	0	3	1	0	2	4	5	2
4	0	2	0	0	1	0	1	0
...	...	...	...	...	...	...	...	...
337332	0	1	0	0	0	0	1	0
337333	0	0	0	0	0	0	0	0
337334	0	1	1	0	0	1	1	0
337335	0	6	0	3	4	2	7	0
337336	0	0	0	0	0	0	0	0

Fig.15: List of Rating

F. CLASSIFICATION OF EMOTIONS:
Emotional interpretation is important to consider the behaviour and fundamental views of individuals. The values of feeling and emotion are identical. In the summary text, we have paired every term with a lexicon document unicorn and increased the number of sentiments for each analysis.[11]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904032 |

	Drug	Condition	Reviews	Rating	Website	anger	anticipation	disgust	fear	joy	negative	positive	sadness	surprise	trust
232363	adderall	adhd	add with mild ocd as well. the med is very eff...	8.0	webmd	1	0	1	1	0	3	1	1	0	1
188173	generess fe	birth control	after reading everyone's horrifying posts i wa...	9.0	webmd	1	6	0	1	1	1	4	0	3	3
14335	etonogestrel	birth control	"i had the implanon after i had my son. i was ...	2.0	drugs.com	0	0	0	0	0	0	0	0	2	0
198901	benzonatate	cough	i have taken the meds for one day and feeling ...	10.0	webmd	0	0	1	0	0	1	0	0	0	0
92841	mucinex d	cough and nasal congestion	"twice this year i have had a cold and used th...	8.0	drugs.com	1	2	0	1	0	2	3	0	0	3

Fig. 16: Classification of Emotions

IX. CONCLUSION

We decided to use opinion mining to extract valuable assumptions through our information that would help consumers, clinics, and practitioners by collecting drug reviews. Utilizing VADER and LSTM framework rating prediction, we were able to identify drugs based on the emotion ranking. We have looked at the 8 feelings that were associated with the use[1].The TF-IDF method produces the highest estimates, with such precision of 83 percent, surpassing baseline simulations. As part of text preprocessing, we used NLP BoW models and various tokenizers to prepare our statistical models. Opinion mining is a branch of research that focuses on extracting information again from the viewpoints of web users. Developing and executing opinion mining techniques is complicated and challenging since the web dataset shows a vast lot of information.

REFERENCES

1. Cabral, Diana &Prudêncio, Ricardo. (2017). Aspect-Based Opinion Mining in Drug Reviews. 815-827. 10.1007/978-3-319-65340-2_66.
2. Vijayaraghavan, Sairamvinay&Basu, Debraj. (2020). Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms.
3. Kaggle https://www.kaggle.com/jessicali9530/kuc-hackathon-winter2018
4.UCI Library Drug Review Dataset: https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Dr ugs.com%29
5. Drug names and their description https://www.drugs.com/ drug_information.html
6. S. M. Jiménez-Zafra, M. T. Martín-Valdivia, M. D. Molina-González, and L. A. Ureña-López, ''How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain,'' Artif. Intell. Med., vol. 93, pp. 50–57, Jan. 2019.
7. F. Gräßer and S. Kallumadi. (Apr. 2018). Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning. [Online].
8. G, Vinodhini&Chandrasekaran, Dr. (2017). Patient opinion mining to analyze drugs satisfaction using supervised learning. Journal of Applied Research and Technology. 15. 10.1016/j.jart.2017.02.005.
9. Sharif, H., Zaffar, F., Abbasi, A., &Zimbra, D. (2014). Detecting Adverse Drug Reactions Using a Sentiment Classification Framework.
10. IoannisKorkontzelos, AzadehNikfarjam, Matthew Shardlow, AbeedSarker, Sophia Ananiadou, and Graciela H. Gonzalez. 2016. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. <i>J. of Biomedical Informatics</i> 62, C (August 2016), 148–158. DOI:https://doi.org/10.1016/j.jbi.2016.06.007
11. Noferesti, S., &Shamsfard, M. (2015). Resource Construction and Evaluation for Indirect Opinion Mining of Drug Reviews. *PLoS ONE, 10.*
12. Na, Jin-Cheon&Kyaing, Wai&Khoo, Christopher & Foo, Schubert & Chang, Yun-Ke&Theng, Yin-Leng. (2012). Sentiment Classification of Drug Reviews Using a Rule-Based Linguistic Approach. 189-198. 10.1007/978-3-642-34752-8_25.
13. Y. Han, M. Liu and W. Jing, "Aspect-Level Drug Reviews Sentiment Analysis Based on Double BiGRU and Knowledge Transfer," in IEEE Access, vol. 8, pp. 21314-21325, 2020, doi: 10.1109/ACCESS.2020.2969473.