Sentiment Analysis on Big Data Using Machine Learning Algorithms

Nikhat Parveen[1], G. Bindu Madhavi[2] , Lakshmi Ramani Burra[3] , Vidyullatha Pellakuri[4], Haran Pellakuri[5]

[1,4]Associate Professor, Department of CSE, KLEF, Vaddeswarm, Guntur(D.t), AP.
[2]Assistant Professor, CSE, Anurag University, Hyderabad, Telangana,
[3]Asst. Professor, Dept. of CSE, PVP Siddhartha Institute of Technology, Kanuru, Vijayawada.
[5]Assistant Professor, Department of CSE, KLEF, Vaddeswarm, Guntur(D.t), AP
[1]nikhat0891@gmail.com, [2]gbindu172@gmail.com, [3]ramanimythili@gmail.com, [4] latha22pellakuri@gmail.com,
[5]haranbabu@kluniversity.in

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

ABSTRACT : Data Analysis which means the data which is obtained that is converted into useful information. Data analysis begins with data through its raw state and turns it into a format that is more readable in the form of graphs, records, plots, etc. allowing it the state and meaning to be understood. After the data is obtained, it reaches the stage of data planning. The level at which raw data is cleaned up and prepared for the next stage of data processing is data planning, also referred to as "pre-processing". Raw data were diligently reviewed for any faults during planning. This move are meant to remove bad data. Therefore the clean data is entered into it and converted into a language it can comprehend. The first step in which raw data starts to take the form of functional information is data entry. The data inserted into the machine in the previous stage is actually analyses for analysis during this step. Machine learning algorithms are used to process the process, but the process itself can differ slightly depending on the data source being processed. The step of output / interpretation is the stage at which data will ultimately be used by non-data researchers. It's accessible and translated.

KEYWORDS: Data analysis, kaggle, ggplot, sentiment analysis, R-studio.

1. INTRODUCTION:

Data analysis happens when data is gathered and converted into useful information. Typically carried out by a data scientist or data science team, it is crucial that data analysis is performed properly in order not to have a detrimental effect on the final result or data production. Data analysis begins with data in its raw state and translates it into a more readable format (graphs, records, etc.), giving it the required form and meaning to be processed by machines and used across an enterprise by employees. The first step in data analysis is to gather information. Data, like data lakes and data centers, is drawn from available sources. It is critical that the available sources of information are accurate and well-built such that perhaps the information gathered (and subsequently included as information) is of the finest quality available. If the data is compiled, it then enters the stage of data processing. The step in which raw data is cleaned up and prepared for the next level of data processing is the preparation of data, also referred to as 'pre-processing'. Raw information is diligently checked for any mistakes during planning[1,2].The purpose of this process are to delete faulty information. Then the clean data is entered into and converted into a format that can be understood. Data entry is the first step in which original data appears to take the form of functional information. The data inserted into the device in the previous stage is then interpreted for analysis during this stage. Machine learning algorithms are used for analysis, but the process itself can differ somewhat depending on source and intended source of the data being processed[3,4]. The level of output / interpretation is the level over which information may ultimately be used by non-data researchers. It's translated and it's legible. Storage is the next step of data analysis. It is then saved for future use until all of the data is analyses. While certain data can be directly put to use, most of it may later have a function. In this paper the movie lens dataset is used to evaluate the user's movie review and generate ggplot on movies throughout the year and figure out the details on how many same genre type movies are released every year and create a contingency table of a certain genre Action. In education, science, and business, the MovieLens datasets are used widely. They are distributed hundreds of thousands of times per year, reflecting their use in novels, conventional and online classes, and applications for popular press programming[5]. In the MovieLens film recommendation scheme, an influential research forum that has hosted several studys since its inception in 1997, these datasets are a result of participant participation. The history of MovieLens and the datasets of MovieLens is documented in this paper. To identify features that influences the ratings of every given movie and evaluates the ratings of the movie. The key challenges of this study are to remove redundant or obsolete data. The quality of findings must be increased by using effective data processing methods USING R Studio 3.1.In social media tracking, sentiment analysis is highly helpful as it helps one to obtain an understanding of the broader public opinion behind such issues[6,7,8]. Owing to real-time tracking features, social media monitoring applications such as Brand watch Analytics make the process smoother and clearer than ever

before.Sentimental study of film feedback is effective in predicting whether or not the film has exceeded the target audience's needs. Such reviews are compiled and formatted from the movie review websites and cleaned and displayed in a csv file. In this research, the dataset used is movie reviews.csv. Six columns are completely included in the dataset (fold id, cv tag, html id, sent id, text and tag). The tag variable says whether or not the comment is favorable by defining the strings for favorable and negative feedback as 'pos' and 'neg' respectively. Text data is evaluated using the Sentiments module in R. This software provides two sentiment-by()and sentiment ()methods for parameter-provided text analysis. The parameters will accept string vectors. The sentiment by feature has a sentiment value of -1 to 1, including 0.0. This role provides the total sentiment meaning, independent of the number of sentences in the text, for the given text [9]. The other sentiment feature provides the sentiment meaning in the given text for each sentence. Sentiment-by is used in this study on a dataset containing approximately 68 K rows. The feedback of any research on the model in this suggested method indicates better performance. This method takes variables / attributes and provides resulting graphs (i.e.) plots of film lens and genre data. And the comprehensive details on Ranking Etc. would be easily available to us.

2. METHEDOLOGY

To collect the data and implement Data Processing, meaning the collected data that is converted into functional information. The processing of data begins with data in its raw state and translates it into a more readable format (graphs, records, etc.), allowing it the state and meaning to be understood. As data is obtained and converted into meaningful content, data processing happens. Typically carried out by a data scientist or a data science team, it is crucial that data analysis is performed properly in order not to have a negative effect on the final result or data production [10, 11, and 13]. In this study, we will conduct data processing and analysis on Movie Lens Here, in order to analyze the data. In this suggested method, the feedback of some analysis on the model produces better performance. This system takes variables / attributes and offers resulting graphs (i.e.) charts of film lens and genre data. And we can easily see the basic details about Scores Etc. Sentiment analysis on film review is the method of evaluating the emotional tone used to develop an appreciation of the attitudes, beliefs and viewpoints behind a set of terms, the feelings reflected in an online guide [12].The nostalgic study of the film is confined to evaluating the movie ratings from the approved websites in order to achieve quality and true movie data. The cinematography and direction trend can be strengthened by studying the movie review data to draw big viewers for the coming movies [14, 15].This allows the organization to expand as well as the viewers to enjoy watching the movie. The "sentimentr" kit included in this study is. The features included in the bundle are used to evaluate the vector sentiment of the strings given to it.

Pseudo code
1. Start
2In a vector, read a dataset
3. Turn it into a Tibble
4. Convert Categorical Knowledge into Variables
5. Retrieve the area of text from the data frame
6. Apply the sentiment-by attribute to the text
7. Storing the results above in a vector
8. For each value val in results vector
a. Assign as 'pos' if val>= 0
b. Assign Else as "neg"
9. Observe the variations between the original and the expected values
10. Stop.

3. EXPERIMENTAL RESULTS:

The experimental results are shown in the following figures from 1 to 5 are the executed results using R studio. The figure 1 shows the number of users reviews given and figure 2 shows the description about movies with generic names. The figure 3,4 and 5 shows the results of ggplot which is plotted the movie data per year and analyzed to find the polarity shows wordcloud graph by describing the sentiment analysis.

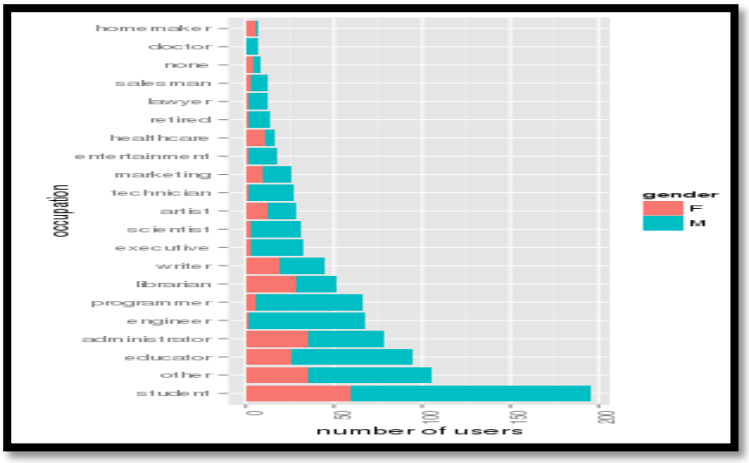
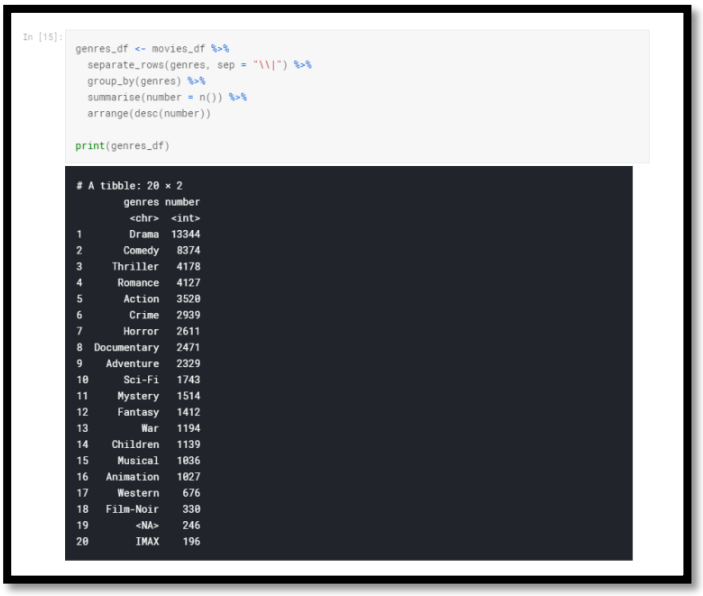


Figure1: different type users Reviews



```
genres_df <- movies_df %>%
  separate_rows(genres, sep = "\\|") %>%
  group_by(genres) %>%
  summarise(number = n()) %>%
  arrange(desc(number))

print(genres_df)
```

```
# A tibble: 20 × 2
       genres number
        <chr>  <int>
1       Drama  13344
2      Comedy   8374
3    Thriller   4178
4     Romance   4127
5      Action   3520
6       Crime   2939
7      Horror   2611
8  Documentary   2471
9   Adventure   2329
10      Sci-Fi   1743
11     Mystery   1514
12     Fantasy   1412
13         War   1194
14    Children   1139
15     Musical   1036
16   Animation   1027
17     Western    676
18   Film-Noir    330
19        <NA>    246
20        IMAX    196
```

Figure2: description about genres

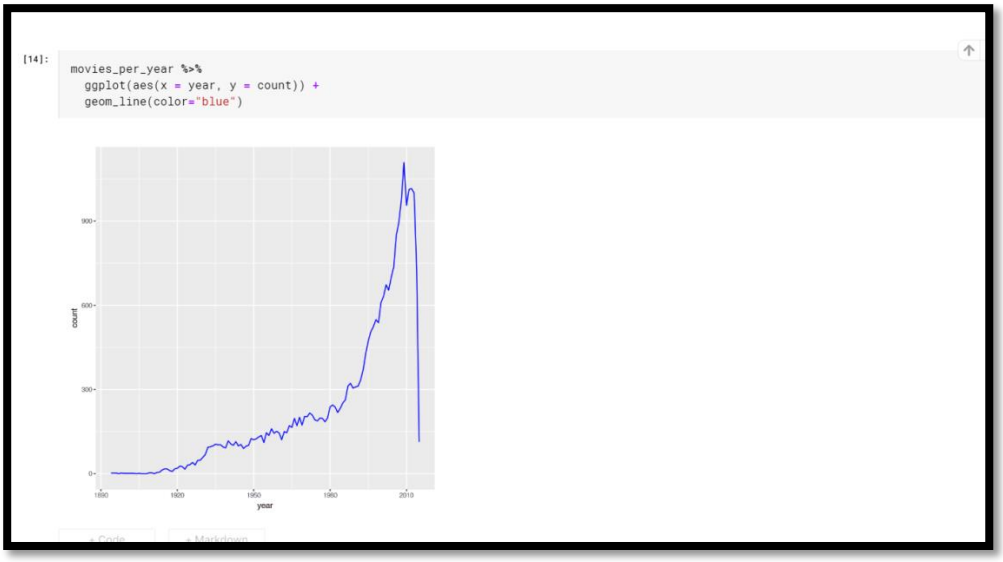


Figure3: ggplot using R showing movies per year

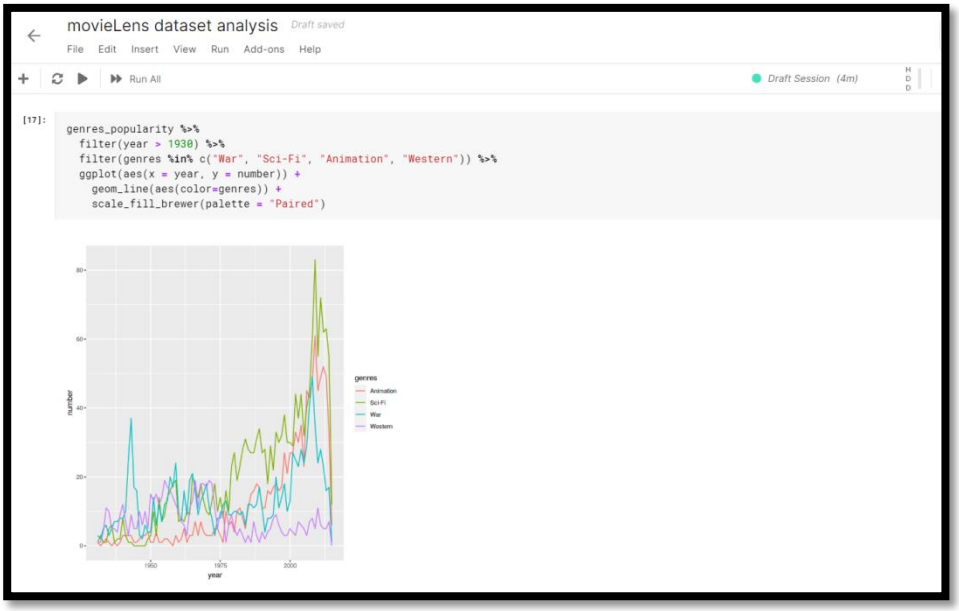


Figure4: movieLens dataset Analysis

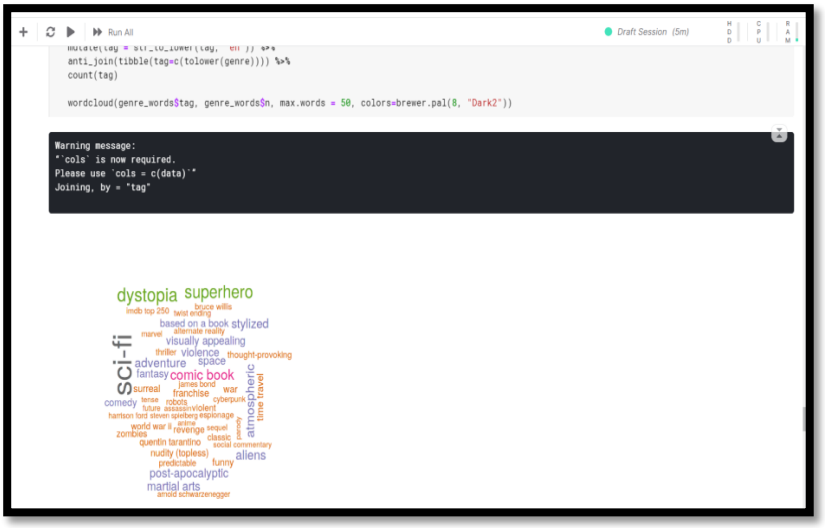


Figure 5: wordcloud showing movie genres

4. CONCLUSION

Many insightful insights into the movie industry were provided by analyzing the movieLens dataset. While it is used mostly for recommendation systems, we were also able to extract some data patterns. The dataset could easily be interpreted to include still more insightful insights using site crawling tools. Overall, analyzing it was a fascinating dataset that allowed many more fascinating R packages & features to be used. This study thus addresses various genres of film lens examination and conducts interpretation and data collection and analysis on the dataset from the Kaggle platform.

References

A. Anila, M., & Pradeepini, G. 2017. Study of prediction algorithms for selecting appropriate classifier in machine learning. Journal of Advanced Research in Dynamical and Control Systems, 9Special Issue 18, 257-268

B. Razia, S., Narasingarao, M. R., & Bojja, P. 2017. Development and analysis of support vector machine techniques for early prediction of breast cancer and thyroid. Journal of Advanced Research in Dynamical and Control Systems, 9Special Issue 6, 869-878.

C. Changala, R., & Rajeswara Rao, D. 2017. A survey on development of pattern evolving model for discovery of patterns in text mining using data mining techniques. Journal of Theoretical and Applied Information Technology, 95(16), 3974-3981

D. Pavan Kumar, N. V. S., & Rajasekhara Rao, K. (2017). Mining negative and positive itemset in parallel and distributed data bases using vertical format. Journal of Advanced Research in Dynamical and Control Systems, 2017

E. Uma Ramya, V., & Thirupathi Rao, K. (2018). Sentiment analysis of movie review using machine learning techniques. International Journal of Engineering and Technology(UAE), 7, 676-681.

F. Anjali Devi, S., Sapkota, P., & Obulesh, M. (2018). Sentiment analysis on products using social media. Journal of Advanced Research in Dynamical and Control Systems, 2018, 137-141.

G. Atmakur, V. K., & Siva Kumar, P. (2018). A prototype analysis of machine learning methodologies for sentiment analysis of social networks. International Journal of Engineering and Technology(UAE), 7, 963-967.

H. Bhargava, M. G., & Rao, D. R. (2018). Sentimental analysis on social media data using R programming. International Journal of Engineering and Technology(UAE), 7(2), 80-84. doi:10.14419/ijet.v7i2.31.13402

I. Vidyullatha pellakuri, "Performance analysis of machine learning techniques for intrusion detection system", 2019Proceedings – 2019, 5th International Conference on Computing, Communication Control and Automation, ICCUBEA 2019

J. Kiran Jammalamadaka, Nikhat Parveen, 2019, Holistic Research of Software Testing and Challenges, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8, Issue- 6

K. Geetha Reddy Y., Prasanth Y. (2019), 'Maximized composite functions based optimized software reliability growth function for reliability prediction', International Journal of Scientific and Technology Research, 8(12), PP.910-919.

L. Mekala M.S., Viswanathan P., Srinivasu N., Varma G.P.S. (2019), 'Accurate Decision-making System for Mining Environment using Li-Fi 5G Technology over IoT Framework', Proceedings of the 4th International Conference on Contemporary Computing and Informatics, IC3I 2019, (), PP.74-79.

M. Krishna Chaitanya, G., Meka, D. R., Vamsi, V. S., & Ravi Karthik, M. V. S. (2018). A survey on twitter sentimental analysis with machine learning techniques. International Journal of Engineering and Technology(UAE), 7(2.32 Special Issue 32), 462-465.

N. Bommadevara, H. S. A., Sowmya, Y., & Pradeepini, G. (2019). Heart disease prediction using machine learning algorithms. International Journal of Innovative Technology and Exploring Engineering, 8(5), 270-272.

O. Kousar Nikhat, A., & Subrahmanyam, K. (2019). Feature selection, optimization and clustering strategies of text documents. International Journal of Electrical and Computer Engineering, 9(2), 1313-1320. doi:10.11591/ijece.v9i2.pp.1313-1320