Journal of Information Systems Engineering and Management
2025, 10(18s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Association Rule Mining and Information Retrieval Using Stemming and Text Mining Techniques

Ashwini Brahme [1], Salim Shaikh [2], Sunita Lokare [3], Sagar Kulkarni [4], Shivaji Mundhe[5], Amit A Jadhav [6]
Nishant Pachpor [7]

[1]Associate Professor, International Institute of Management Science, Pune, India,
[2]Associate Professor, Dept. of Computer Engineering, Anjuman I Islam's Kalsekar Technical Campus, School of Engineering and Technology, New Panvel, Navi Mumbai,
[3]Dr. D.Y. Patil Ambi, Pune, India
[4]. Mumbai University
[5] Director, International Institute of Management Science, Pune, India,
[6]. Assistant professor, D.Y.Patil University Pune, Ambi, India,
[7]Assistant Professor, International Institute of Management Science, Pune, India,

ARTICLE INFO

Received: 15 Dec 2024

Revised: 26 Jan 2025

Accepted: 15 Feb 2025

ABSTRACT

Heterogeneous, complex and enormous data mining plays significant role in the today's big data scenario all over the globe. The research paper is intended toward the natural language processing, mining of textual data, and pattern discovery through association rule mining. The research is aimed towards mining of digital news of epidemic diseases and generating the hidden patterns from the corpus data. The present study also aimed towards developing knowledge discovery system for healthcare for prediction of epidemic viral diseases and their related measures which will be helpful for the healthcare experts, doctors, and healthcare organizations as well as for governments also to take the precautionary measures. The study deigned for predictive analytics of epidemic diseases and their patterns using association rule mining. The precautionary measures for the healthcare and highly impacted geographical location of widespread diseases are generated through the proposed system.

Keywords: Association Rule mining, Text Mining, Stemming, Predictive Analytics, Knowledge Discovery, epidemic viral diseases, frequent patterns, data mining , Apriori Algorithm , text pre-processing , tokenization's, pattern.

INTRODUCTION

The advancement of health informatics, particularly through electronic health records, has led to an overwhelming influx of data within healthcare organizations. These systems store vast amounts of diverse patient information, including medical history, lab test results, admission details, and personal statistics, which are managed by a range of professionals. As a result, many healthcare organizations and leaders are actively seeking solutions to effectively handle this massive data for improved patient care and treatment. However, analyzing this data presents several challenges, such as correlating patient demographics with critical illnesses, gaining deeper insights into symptoms and their causes, and optimizing treatment strategies.

Knowledge discovery and data mining play a crucial role in addressing these challenges, with applications spanning four key areas: clinical medicine, healthcare policy and planning, public health, and healthcare text mining. Given the complexity of healthcare data—where multiple variables interact in subtle ways—knowledge discovery helps uncover hidden patterns and relationships within extensive databases. This can be particularly beneficial for predicting at-risk patients, diagnosing and treating diseases, managing chronic conditions, and addressing public health concerns. The ability to extract meaningful insights from large datasets ultimately enhances decision-making, promotes early intervention, and improves overall healthcare outcomes.

It is essential to focus on predicting the spread of infectious diseases, as their prevalence has increased significantly and expanded across various geographic regions. This rise can be attributed to factors such as population growth, international travel, global trade, and the transmission of infections and viruses. Understanding these contributing factors is difficult for adopting preventative measures of widespread diseases and unawareness of the immediate medication; also, its consequences resulted in more adverse impacts on public health. Nowadays, lot of information

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

is available about these epidemic (highly spread) disease out-breaks in the form of newspapers, websites/internet, healthcare/life science, publications/reports, presentations and many more. It's very challenging to follow and analyze each and every information channel in its entirety to derive tangible conclusions – say, all news-papers publications.

By reading all information available in newspapers, it is impossible to gather accurate knowledge for predicting high frequency diseases. The below listed vital factors are required to arrive at any kind of swift medical conclusion in deciding further line of treatment (LoT)

- The geographical location – Important in terms of measuring the wide-spread
- Number of Victims & Impacts – To check on the penetration and pin-down the type of infectious virus
- Line of treatment and remedy– To predicate and analyze the possible disease, geographical location and precautionary measures

Hence there is need to study the predictive analytics.

REVIEW OF LITERATURE

A study is focused on association rule mining of e-news using FP growth and Apriori algorithm, the massive data of semi structured and unstructured format is used to mining and get the appropriate knowledge. Almost 80% of data all over the globe is on internet and to take that and mine it with respect to specific domain and to generate the knowledge from this data is the biggest challenging task. Text mining is the best solution to mine the corpus data, process it, and generate the effective knowledge discovery which is fruitful for the decision making. Association rule mining comprised of various steps like pre-processing; it includes text transformation, tokenization, and filtration, stemming and indexing. The researcher has compared the algorithms namely Apriori and FP Growth in terms of text processing and generative effective association rules for better knowledge discovery. [23]

The research paper entitled "A Comparative Analysis of Association Rule Mining Algorithms" comprised of knowledge discovery from database. This research has focused on association rule mining gone of the best data mining technique for pattern generation and frequent item set identification. The present study focused on different algorithms utilized in Association Rule Mining (ARM), specifically AIS, SETM, AprioriTid, AprioriHybrid, FP-Growth, and LogEclate. The evaluation emphasizes critical factors such as accuracy, processing speed, and database support. Each algorithm presents unique execution times, advantages, and limitations. Especially, the complexity of the FP-Growth algorithm can be a drawback, contributing to the overall complexity of the method. [24]

In association rule mining, frequent item sets are identified by scanning and extracting data with lowest support threshold. These item sets are evaluated with lowest confidence level to reveal hidden patterns and valuable insights applicable across various domains. The PPARM algorithm emerges as an optimal solution for safeguarding complex rules while ensuring mining accessibility, particularly leveraging the power of metaheuristic algorithms. This paper highlights future opportunities for innovative approaches and frameworks in privacy-preserving data mining, effectively with real time applications. [25]

This study examines key occupational risk factors impacting miners' health and psychosocial well-being. Association rule mining is vital for meta-analysis of stressors & vulnerability factors linked to miners' health, systematically identifying patterns and their corresponding health outcomes. The different patterns using association rule mining are generated related to health and risk as well as stress parameters of workers working in mines. [26]

Research Gap:

The research gap is identified from text mining in healthcare, knowledge discovery, stemming technique. The previous studies are carried out on few aspects of text mining in various sectors; but text mining of large amount of viral diseases study for effective pattern discovery has not been carried out. Also knowledge discovery, predictive analysis and remedial measures suggestions through the system is not carried out to solve the analyzed problems therefore researcher had made an attempt to resolve this gap.

MOTIVATION OF THE STUDY

Healthcare is a broad domain where significant efforts are dedicated to knowledge discovery. Manually analyzing vast amounts of data and extracting meaningful insights is impractical due to the abundance of information available online for the same text mining is helpful. Researchers primarily focus on identifying common patterns, correlation of item sets related to viral diseases to facilitate faster and more informed decision-making. To address challenges in text mining, pattern discovery, information retrieval, stemming, and knowledge extraction, a system was designed, developed, and implemented.[28] The core objective intended towards prediction of high-impact viral infectious diseases and geographical spread, enabling healthcare organizations, medical professionals, and local authorities to take prompt preventive measures and mitigate the risk of widespread health crises.

OBJECTIVES

1. To study the epidemic highly frequent diseases
2. To understand the need of predictive analytics for infectious diseases prediction.
3. To evolve and implement Prep-NPS stemming algorithm.
4. To apply association rule mining from the selected dataset of viral diseases.
5. to design and implement the predictive analytics system for healthcare

RESEARCH DESIGN

The current study is particularized as:

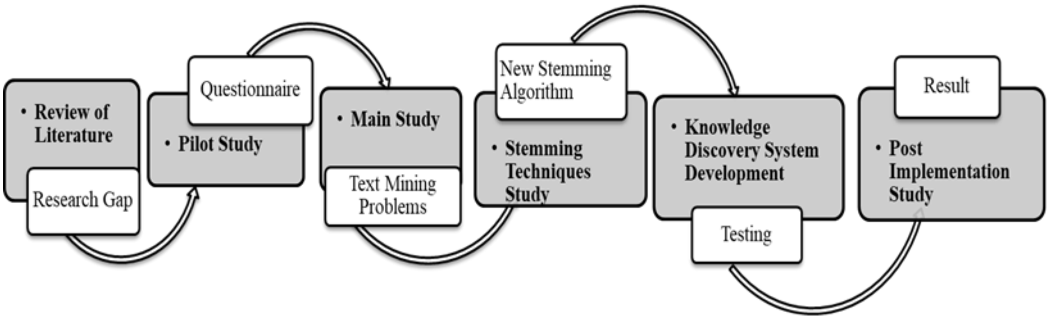


Figure 1: Research Design compiled by researcher

The introductory things are comprised of text mining , knowledge discovery , applications of knowledge discovery , challenges of heterogonous, complex data and generating the effective patterns from them .The research is focused on healthcare , highly spread diseases, its patterns, and need of predictive analysis of epidemic diseases . The research then focuses on reviewing literature of association mining, KDD, text and data mining, information retrieval from textual data, stemming techniques and decision-making related to viral diseases and their outbreaks. The research done the said area and research gaps are identified for further scope and innovations.

The scope of research is considered with respect to conceptual scope, geographical scope and technical scope. The data of highly spread viral diseases along with various parameters of diseases is collected through first hand questionnaire, interview technique the data collection is carried out. Classification of online news of highly frequent selected viral diseases is carried out and text mining is carried out on the online news dataset of selected diseases.

Stemming is carried out on selected stemming techniques algorithms, a new stemming algorithm is introduced namely Prep-NPS. The results generated using this algorithm are considered for association rule mining using association rule induction tool.

The association rules are considered for knowledge discovery and predictions of diseases and their demographics. The developed application tested and found to be satisfactory through the expert's decision. The generated result are considered for decision making in healthcare and related stakeholders.

DESIGNED KNOWLEDGE DISCOVERY SYSTEM

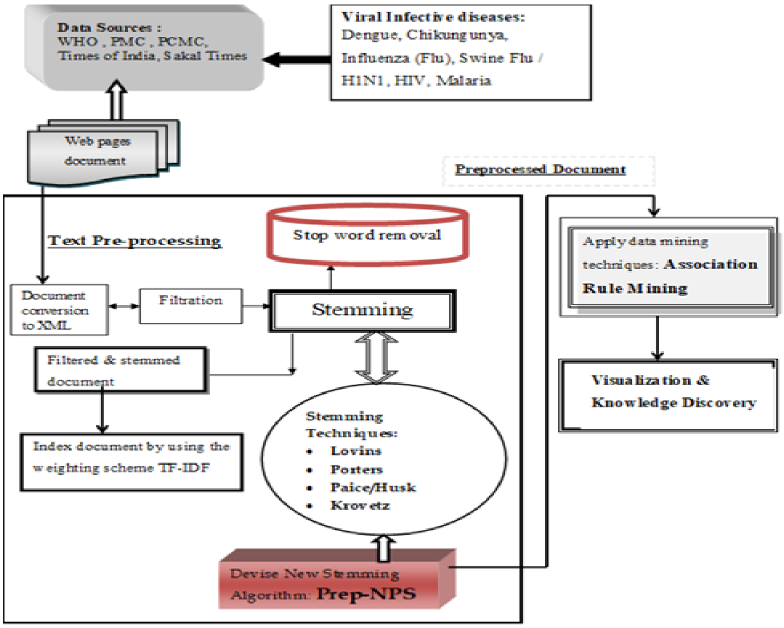

Data Sources :
WHO , PMC , PCMC,
Times of India, Sakal Times
Viral Infective diseases:
Dengue, Chikungunya,
Influenza (Flu), Swine Flu /
H1N1, HIV, Malaria
Web pages document
Preprocessed Document
Text Pre-processing
Stop word removal
Document conversion to XML
Filtration
Stemming
Filtered & stemmed document
Index document by using the weighting scheme TF-IDF
Stemming Techniques:
• Lovins
• Porters
• Paice/Husk
• Krovetz
Devise New Stemming Algorithm: Prep-NPS
Apply data mining techniques: Association Rule Mining
Visualization & Knowledge Discovery

Figure 2: Knowledge discovery system for Healthcare

Source: Designed by researcher

Stemming: The information retrieval of textual data need stemming therefore researcher has applied various stemming techniques to get the proper root word and use it for association mining. The porters stemming has limitations while generating the stems as it works on prefix stripping and suffix stripping. To improve the performance of Porters stemming the researcher has designed new algorithm Prep-NPS for accurate result and for correct association rule mining.[27]

The accuracy and errors in stemming for Prep-NPS, Porter's, Lovins, Paice-Husk, and Krovetz algorithms are summarized in the following table. Among 30,520 tokens, Prep-NPS produced 28,319 correct and 2,201 incorrect stems, while Porter's algorithm resulted in 20,341 correct and 10,179 incorrect stems. Lovins generated 13,590 correct and 16,930 incorrect stems, PaiceHusk yielded 12,598 and 17,922 correct and incorrect stems consequently while as Krovetz produced 20,088 and 10,432 correct and incorrect stems sequentially.

Prep-NPS developed stemming algorithm achieves an accuracy 93 % %, with only 7.21% incorrect stems. In comparison, Porter's algorithm produces 67% and 34 % correct &incorrect stems. 45% and 55% correct & incorrect stems resulted through Lovins algorithm , Paice-Husk generated 41% , 59 % correct & incorrect stems consequently while as Krovetz resulted 66% correct and 34% incorrect stems. Hence, Prep-NPS stemming algorithm outperforms the other methods, providing the highest accuracy and the maximum number of correct stems compared to Porter's, Krovetz, Lovins, and Paice-Husk algorithms.

Association rule mining: The present study utilized the open-source Association Rule Induction Tool, developed by Christian Berget, which offers a user-friendly graphical interface based on the Apriori algorithm for identifying association rules. For this analysis, the researcher selected the top 1,000 distinct tokens with the highest TF-IDF scores. The large number of association rules were generated by an association Rule induction tool. A total of 41,708 association rules were derived from 1,000 distinct tokens within 6.673 seconds. The generated rules are applied in the healthcare knowledge discovery for knowledge generation and decision making.

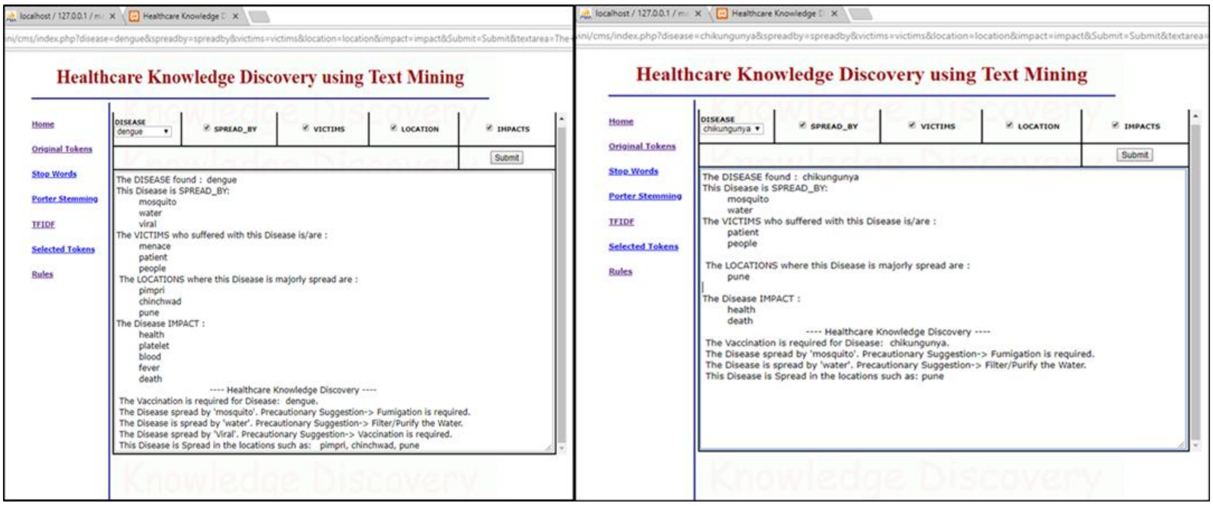
Table 1: Association rules of viral diseases

Sr. no.	**Association Rule**	**Support**	**confidence**	**Interpretation**
Rule 5	{swab}<- { tamiflu, swine } or { tamiflu, swine } <- { swab}	10.8	76.5%	there is relation of swab is taken/ tested for related with

				swine flu diseases and need of Tamiflu tablet
Rule 178	{pmc } <- { tamiflu, flu } or { tamiflu, flu } <- { pmc }	10.8	94.1 %	This indicates that the PMC area has a higher number of flu patients, highlighting the need for Tamiflu medication.
Rule 395	{platelet} <- {count, dengue} or {count, dengue} <-{platelet}	11.5	72.2 %	The rule suggests that if a patient is suspected of having dengue, their platelet count should be monitored.
Rule 1845	dengue} <- {mosquitoes, breede} or {mosquitoe, breede} <- {dengue}	12.1	68.4%	It outcomes dengue diseases take places due to mosquitoes breede.

Source: Compiled by researcher

Knowledge Discovery through Association rules: The generated association rules taken for knowledge discovery; the system is trained for the following diseases with various parameters namely disease, spread by, victims, impact and location. The designed system selects diseases and displays the patterns in terms of SPREAD-BY, VICTIM, LOCATIONS and IMPACT and generates the knowledge in terms of vaccination is required, precautionary measures and suggestions. Also the system predicts highly spread diseases geographical locations so that the healthcare stakeholders can take remedial actions and have control of epidemic diseases represented in the above GUI for Dengue Diseases, Chikungunya.


localhost / 127.0.0.1 / m
Healthcare Knowledge D
ni/cms/index.php?disease=dengue&spreadby=spreadby&victims=victims&location=location&impact=impact&Submit=Submit&textarea=The
Healthcare Knowledge Discovery using Text Mining
Home
Original Tokens
Stop Words
Porter Stemming
TFIDF
Selected Tokens
Rules
DISEASE
dengue
SPREAD_BY
VICTIMS
LOCATION
IMPACTS
Submit
Knowledge Discovery

```
The DISEASE found : dengue
This Disease is SPREAD_BY:
        mosquito
        water
        viral
The VICTIMS who suffered with this Disease is/are :
        menace
        patient
        people
 The LOCATIONS where this Disease is majorly spread are :
        pimpri
        chinchwad
        pune
The Disease IMPACT :
        health
        platelet
        blood
        fever
        death
                    ---- Healthcare Knowledge Discovery ----
 The Vaccination is required for Disease:  dengue.
 The Disease spread by 'mosquito'. Precautionary Suggestion-> Fumigation is required.
 The Disease is spread by 'water'. Precautionary Suggestion-> Filter/Purify the Water.
 The Disease spread by 'Viral'. Precautionary Suggestion-> Vaccination is required.
 This Disease is Spread in the locations such as:   pimpri, chinchwad, pune
```

localhost / 127.0.0.1 / m
Healthcare Knowledge D
ni/cms/index.php?disease=chikungunya&spreadby=spreadby&victims=victims&location=location&impact=impact&Submit=Submit&textarea=
Healthcare Knowledge Discovery using Text Mining
Home
Original Tokens
Stop Words
Porter Stemming
TFIDF
Selected Tokens
Rules
DISEASE
chikungunya
SPREAD_BY
VICTIMS
LOCATION
IMPACTS
Submit
Knowledge Discovery

```
The DISEASE found : chikungunya
This Disease is SPREAD_BY:
        mosquito
        water
The VICTIMS who suffered with this Disease is/are :
        patient
        people

 The LOCATIONS where this Disease is majorly spread are :
        pune

The Disease IMPACT :
        health
        death
                    ---- Healthcare Knowledge Discovery ----
 The Vaccination is required for Disease:  chikungunya.
 The Disease spread by 'mosquito'. Precautionary Suggestion-> Fumigation is required.
 The Disease is spread by 'water'. Precautionary Suggestion-> Filter/Purify the Water.
 This Disease is Spread in the locations such as: pune
```

Figure 3: Knowledge discovery of viral diseases

The designed system tested using functionality testing; bugs and defects were eliminated to enhance quality of system functional testing was conducted, relevant test cases are written and taken for checking the and evaluating functionality of designed application.

CONCLUSION

An internet contains an overwhelming volume of data, much of which is in unstructured or semi-structured formats, including text, emails, images, graphs, audio, videos, and blogs. This study focuses on text mining news and information related to specific viral infectious diseases such as HIV, Influenza Flue/Flue, Chikungunya, Dengue, Diarrhea, and Swine Flu, considered form various digital sources WHO, English newspapers, online News etc. To enhance the information retrieval and processing the stemming is carried out. The present research focused on stemming algorithms, including Porter's, Paice/Husk, Lovins, and Krovetz, to generate both correct and incorrect stems. Stemming is crucial for effective information retrieval, obtaining root words, association rule mining, and knowledge discovery. To improve accuracy, researcher has developed and implemented a new stemming algorithm called Prep-NPS, which was then compared to the selected algorithms based on the correctness of stemming. The hypothesis was tested and analysed statistically. The present research resulted into effective implementation of stemming algorithm, information retrieval, association rule mining, and knowledge discovery from textual data. The system results in significant prediction of highly spread viral infective diseases and its geographical location which is beneficial for doctors, practitioners, medical professionals, hospitals, dispensaries, healthcare departments, pharmaceutical organizations, test laboratories, Life-science organizations, society, social media and various websites. Therefore, the researcher assures that the outcomes of this study will be valuable and impactful healthcare experts in the line of decision making, preventive measures and making awareness to the society.

REFRENCES

[1] Viral Diseases.

[2] http://www.rightdiagnosis.com/v/viral/intro.htm

[3] Viral Infections. US National Library of Medicines. https://medlineplus.gov/viralinfections.html

[4] Mandal A. (2013). Human Diseases Caused by Viruses.http://www.newsmedical.net/health/Human- Diseases-Caused-by-Viruses.aspx

[5] Dengue Fever News and Research. News Medical Life Sciences. http://www.news medical.net/?tag=/Dengue-Fever.

[6] https://www.healthgrades.com/conditions/ viral-diseases

[7] https://pmc.gov.in/en/health.

[8] https://pmc.gov.in/en/vector-borne- disease-control-0.

[9] [8]. https://www.pcmcindia.gov.in/departments-details.php?Id=16.

[10] http://www.who.int/en/

[11] http://www.statisticshowto.com/probabilit y-and-statistics/t-test/

[12] https://pmc.gov.in/en/vector-borne- disease-control

[13] Kulkarni A., Mundhe S. (2017). A study of Viral Diseases and their Impact on Public Health using Knowledge Based System. BIONANO FRONTIER.10 (2). Page.273

[14] Ojo A., Adeyemo A. (2012). Framework for Knowledge Discovery from Journal Articles UsingText Mining Techniques. African Journal of Computing & ICT, 5(2). www.ajocict.net

[15] Patel M., Patel A., Virparia P. (2013). Rule Based Expert System for Viral Infection Diagnosis. International Journal of Advanced Research in Computer Science and Software Engineering, 3(5).www.ijarcsse.com

[16] K. Prabavathy, P. Sumathi (2013). Text Mining Interpreting Knowledge Discovery from Biomed Articles. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), 1(2), 33- 36

[17] H Shaker. El. S., El-MasriS. ,Riad A., Elmogy M.(2013). Data Mining and Knowledge Discovery: Applications, Techniques, Challenges and Process Models in Healthcare. International Journal of Engineering Research and Applications (IJERA), 3(3),900-906. www.ijera.com

[18] Amin A., Talib R., Raza S. (2014). Extract association rules to minimize the effects of dengue by using a text mining technique. International Journal of Computer Science and Mobile Computing, 3(4), 394-400. www.ijcsmc.com

[19] SHAKIL K., ANIS S., ALAM M. (2013). Dengue Disease Prediction Using Weka Data Mining Tool. Department of Computer Science, Jamia Millia Islamia New Delhi, India

[20] Brahme A., Mundhe S. (2022). To Devise and Implement Effective Stemming Algorithm for Text Mining and NLP for Healthcare Management. Journal of Management and Entrepreneurship (JME) Page no. 347

[21] Brahme A., Mundhe S. (2021). Text Mining: Application of Stemming Algorithms for Information Retrieval in Healthcare with special reference to Viral Infective diseases. Shodhsanchar.11 (41).

[22] https://www.mdpi.com/2071- 1050/13/16/8900

[23] Agrawal and R. Srikanth, “Fast Algorithms for Mining Association Rules,” In Proc. of VLDB ’94, pp. 487-499, Santiago, Chile, Sept. 1994.

[24] R. Agrawal, T. Imielinski, and A. Swami (1993). Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD International Conference on Management of Data - SIGMOD '93. p. 207 Washington, D.C., May 1993.

[25] T. Lakshika, A. Caldera "Association Rules For Knowledge Discovery from E-News Articles: A Review Of Apriori And FP-Growth Algorithms", Advances in Science, Technology and Engineering Systems Journal, vol. 7, no. 5, pp. 178-192 (2022). (https://www.astesj.com/v07/i05/p19/#1639848202927-7c57dd68-5f71)

[26] S. S. Aljehani and Y. A. Alotaibi, "Preserving Privacy in Association Rule Mining Using Metaheuristic-Based Algorithms: A Systematic Literature Review," in IEEE Access, vol. 12, pp. 21217-21236, 2024, doi: 10.1109/ACCESS.2024.3362907.

[27] Zhang, B.; Yin, X.; Guo, Y.; Tong, R. What occupational risk factors significantly affect miners' health: Findings from meta-analysis and association rule mining. J. Saf. Res. 2024, 89, 197–209.

[28] R. Kulkarni and D. S. D. Mundhe, "Data Mining Technique: An Implementation of Association Rule Mining in Healthcare", Int. Adv. Res. J. Sci. Eng. Technol., vol. 4, no. 7, pp. 62-65, 2017.

[29] Kulkarni, A., & Mundhe, S. (2016). A theoretical review on text mining: Tools, techniques, applications and future challenges. Int. J. Innov. Res. Comput. Commun. Eng, 4(11), 19225-19230.

[30] Kulkarni A.R., Mundhe S.D.(2019). An application of porters stemming algorithm for text mining in healthcare. Int. J. Manag. IT Eng., 7 (11), pp. 223-228