wjert, 2020, Vol. 6, Issue 1, 233-248. *Original Article* ISSN 2454-695X

World Journal of Engineering Research and Technology

WJERT

www.wjert.org

SJIF Impact Factor: 5.924

STUDY ON TEXT DATA MINING FOR CARE LIFE LOG USING KEY GRAPH

Devanshi Nayak*[1] and Prof. S. R. Yadav[2]

[1]M.Tech. Scholar, CSE, MITS Bhopal.

[2]A. P., CSE, MITS Bhopal.

Article Received on 01/12/2019 **Article Revised on 22/12/2019** **Article Accepted on 12/01/2020**

***Corresponding Author**
Devanshi Nayak
M.Tech. Scholar, CSE,
MITS Bhopal.

ABSTRACT

Text data mining is a process of extracting interesting and nontrivial patterns from huge amount of text documents. There exist different techniques and tools to mine the text and discover valuable information for future prediction and decision making process. The selection of right and appropriate text mining technique helps to enhance the speed and decreases the time and effort required to extract valuable information. Care Life Log is used to integrate and analyze the level of care required. There are five levels of care, with Level 1 vocabulary including recreation, toilet, morning, afternoon, etc. The level of care gradually increases from Level 1 to Level 5, which has vocabulary that includes tube, danger, treatment, removal, and discovery. The higher the level, the worse the health condition and therefore the greater care required. These levels allow for a clear analysis of a patient's condition. This analysis has led to an improvement in Quality of Life as well as a decrease in mismatches between the level of care required for patients and the level of care given by care takers. Rapid progress in digital data acquisition techniques have led to huge volume of data. More than 80 percent of today's data is composed of unstructured or semi-structured data. The discovery of appropriate patterns and trends to analyze the text documents from massive volume of data is a big issue. A text data mining technique identified the relations between feature vocabularies seen in past in-patient records accumulated on the University of Miyazaki Hospital's Electronic Medical Record, and extractions were made. The qualitative analysis result of in-patient nursing records used a text data mining technique to achieve the initial goal: a visual record of such information. The analysis discovered vocabularies relating to proper treatment methods and concisely

summarized their extracts from in-patient nursing records. Important vocabularies that characterize each nursing record were also revealed. The results of this research will contribute to nursing work evaluation and education.

KEYWORDS: Text Data Mining, Care Life Log, Nursing Records, Qualitative Analysis, Data Acquisition, Electronic Medical Records.

1. INTRODUCTION

Text mining is a process to extract interesting and significant patterns to explore knowledge from textual data sources.[3] Text mining is a multi-disciplinary field based on information retrieval, data mining, machine learning, statistics, and computational linguistics.[3] Figure 1 shows the Venn diagram of text mining and its interaction with other fields.

Several text mining techniques like summarization, classification, clustering etc., can be applied to extract knowledge. Text mining deals with natural language text which is stored in semi-structured and unstructured format.[4] Text mining techniques are continuously applied in industry, academia, web applications, internet and other fields.[5] Application areas.

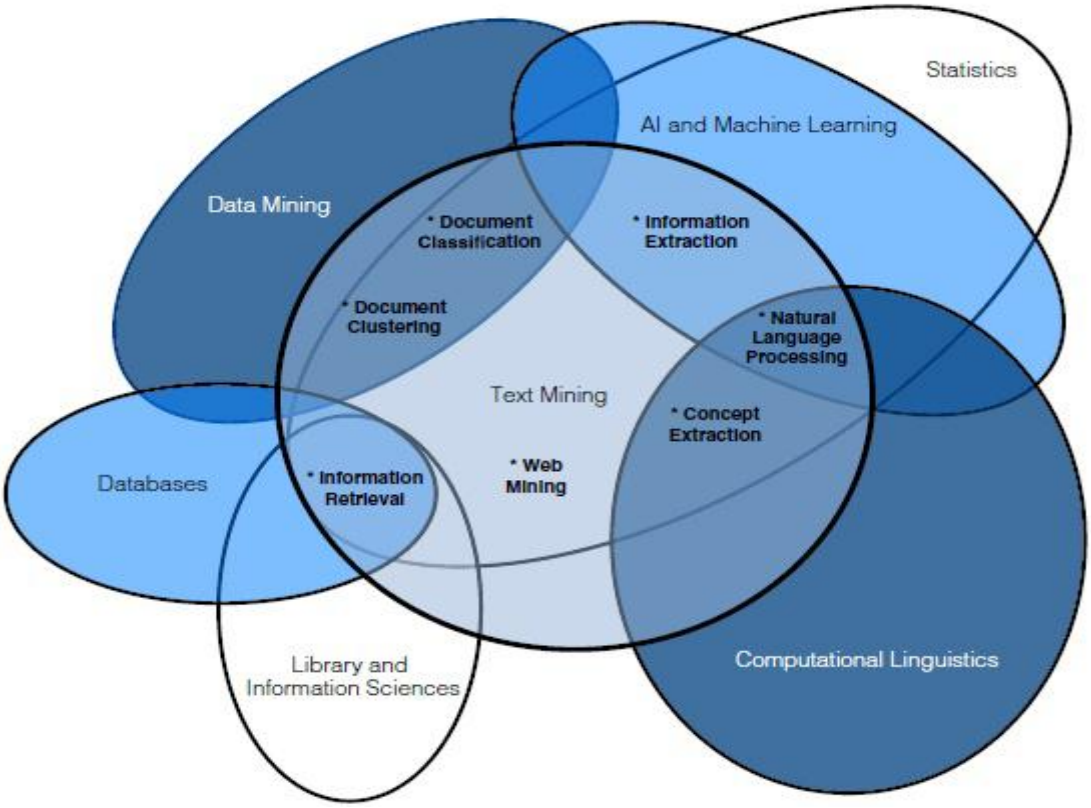


Fig. 1: Venn diagram of text mining interaction with other fields.

like search engines, customer relationship management system, filter emails, product suggestion analysis, fraud detection, and social media analytics use text mining for opinion mining, feature extraction, sentiment, predictive, and trend analysis.[6] Generic process of text mining performs the following steps (Figure 2).

_ Collecting unstructured data from different sources available in different file formats such as plain text, web pages, pdf files etc.

_ Pre-processing and cleansing operations are performed to detect and remove anomalies. Cleansing process make sure to capture the real essence of text available and is performed to remove stop words stemming (process of identifying the root of certain word) and indexing the data.[7]

_ Processing and controlling operations are applied to audit and further clean the data set by automatic

processing.

_ Pattern analysis is implemented by Management Information System (MIS).

_ Information processed in the above steps are used to extract valuable and relevant information for effective and timely decision making and trend analysis.[8]

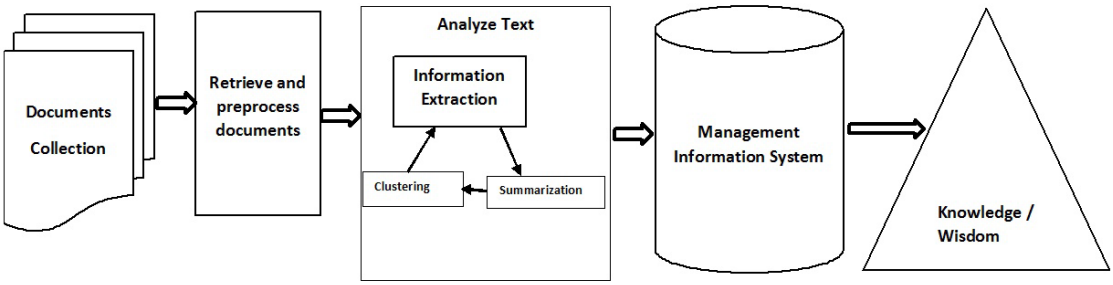


Fig. 2: Text mining process.

Extraction of valuable information from a corpus of different document is a tedious and tiresome task. The selection of appropriate technique for mining text reduce the time and effort to find the relevant patterns for analysis and decision making.

1. LITERATURE REVIEW

Kenji Araki and Muneou Suzuki, IAENG International Journal of Computer Science· August 2011, 38:3, IJCS_38_3_05, Text Data Mining of In-patient Nursing Records Within Electronic Medical Records Using KeyGraph. The analysis discovered vocabularies relating

to proper treatment methods and concisely summarized their extracts from in-patient nursing records.

S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11 303–11 311, 2012. described that gathering, extracting, pre-processing, text transformation, feature extraction, pattern selection, and evaluation steps are part of text mining process. In addition, different widely used text mining techniques, i.e., clustering, categorization, decision tree categorization, and their application in diverse fields are surveyed.

N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE transactions on knowledge and data engineering, vol. 24, no. 1, pp. 30–44, 2012. Highlighted the issues in text mining applications and techniques. They discussed that dealing with unstructured text is difficult as compared to structured or tabular data using traditional mining tools and techniques. They have shown the applications of text mining process in bioinformatics, business intelligence and national security system. Natural language processing and entity recognition techniques has reduced the issues that occur during text mining process. However, there exist issues which need attention.

A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravi˘cius, and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," Journal of biomedical semantics, vol. 5, no. 1, p. 1, 2014. explored MEDLINE biomedical database by integrating a framework for named entity recognition, classification of text, hypothesis generation and testing, relationship and synonym extraction, extract abbreviations. This new framework helps to eliminate unnecessary details and extract valuable information.

B. Laxman and D. Sujatha, "Improved method for pattern discovery in text mining," International Journal of Research in Engineering and Technology, vol. 2, no. 1, pp. 2321–2328, 2013. analyzed the text using text mining patterns and showed term based approaches cannot analyze synonyms and polysemy properly. Moreover, a prototype model was designed for specification of patterns in terms of assigning weight according to their distribution. This approach helps to enhance the efficiency of text mining process.

Muneo Kushima, Kenji Araki, Tomoyoshi Yamazaki, Sanae Araki, Taisuke Ogawa, Noboru Sonehara, Proceedings of the International Multi Conference of Engineers and Computer

Scientists 2017 Vol I, IMECS 2017, March 15 - 17, 2017, Hong Kong. Text Data Mining of Care Life Log by the Level of Care Required Using KeyGraph. This analysis has led to an improvement in Quality of Life as well as a decrease in mismatches between the level of care required for patients and the level of care given by caretakers.

Muneo Kushima, Tomoyoshi Yamazaki and Kenji Araki, Proceedings of the International Multi Conference of Engineers and Computer Scientists 2019 IMECS 2019, March 13-15, 2019, Hong Kong. Text Data Mining of the Nursing Care Life Log from Electronic Medical Record. The analysis results have potential to clarify the work content of care workers. As the nursing field requires efficiency in health care services, improvement and continuous data collection are important for the long-term building of health care services as well as large-scale data collection.

2. METHODOLOGY

Key Graph

Key Graph consists of three major components derived from construction metaphors. Each component is described as follows:

1) Foundations: sub-graphs of highly associated and frequent terms that represent basic concepts in the data. A foundation is defined as a cluster that consists of black nodes linked by solid lines. The foundations are underlying common contexts because they are formed by a set of items that frequently co-occur in the data set.

2) Roofs: terms that is highly associated with foundations.

3) Columns: associations between foundations and roofs that are used for extracting keywords, i.e., the main concepts in the data. A column is a dotted line that connects foundations.

Example of Key Graph Performance

Figure 3 shows an example when it is applied to text data.

1. Black nodes indicate items that frequently occur in a data set.
2. White nodes indicate the items that occur less frequently overall but frequently occur with black nodes in a data set.
3. Double-circled nodes indicate items whose co-occurrence frequency with black nodes is especially high. Double-circled nodes are considered keywords.
4. Links indicate that the connected item pair frequently co-occurs in a data set.

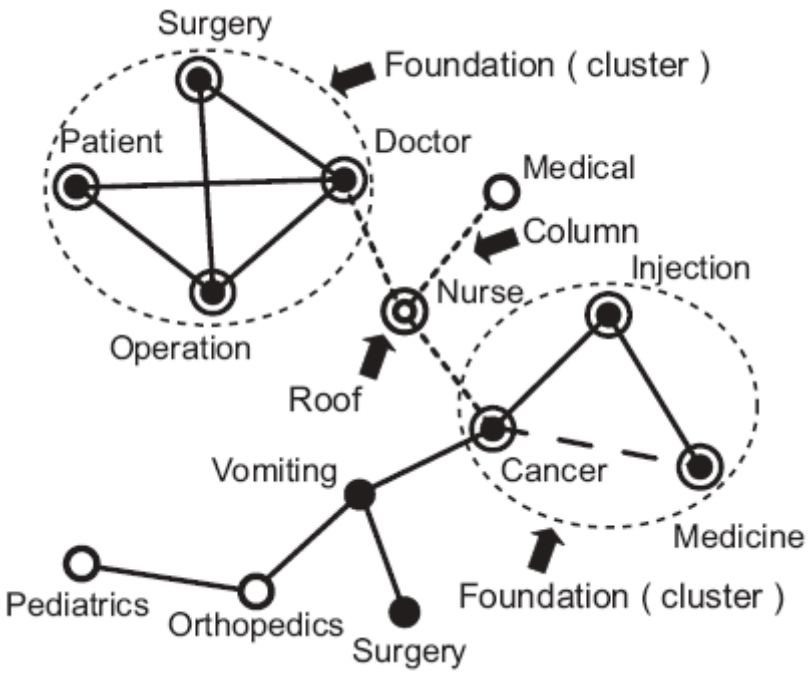
5. Solid lines form a foundation, which dotted lines connect. Foundations, which are circles of dotted lines, are obtained from the text data.

●: High frequency terms

○: Connect high frequency "foundations"

◙: Keywords of ●

©: Keywords of ○



"Foundation (Cluster)"

: Mass of black nodes and links

"Roof": Connect with the "foundation"

"Column": ----------

: Significant "roofs" between "foundations"

Solid lines (-): Strongly correlated terms of "foundations"

----------: Strongly correlated terms

Fig. 3: Key Graph example when applied to text data.

Key Graph Algorithm

Key Graph was originally an algorithm for extracting assertions based on the co-occurrence graph of terms from text data.

Its process consists of four phases:

1) Document preparation: Before processing document D, stop words [29] with little meaning are discarded from, the words in D are stemmed [30], and the phrases in D are specified [31]. Hereafter, a term means a word or a phrase n processed D.
2) Extracting foundations: Graph G for document D is made of nodes representing terms and links representing the co-occurrence (term-pairs that frequently occur in the same sentences throughout D). Nodes and links in G are defined as follows:

Nodes: Nodes in G represent high-frequency terms in D because they might appear frequently for expressing typical, basic concepts in the domain. High-frequency terms are a set of terms above the 30th highest frequency. This set is denoted by HF.

Links: nodes in HF are linked if the association between the corresponding terms is strong. The association of terms between wi and wj in D is defined as assoc (wi, wj) = $\sum_{S\epsilon D} \min(|wi|s, |wj\ |s),$ (1)

Where |x|s denotes the count of x in sentence s. Pairs of high-frequency terms in HF are sorted by assoc and the pairs above the (number of nodes in G) -1th tightest association are represented in G by links between nodes.

Extracting columns: The probability that term w appears is defined as key (w), and key (w) is defined by

Key (w) = 1 −Πg⊂G [1 – based (w, g) / neighbors (g)], (2)

Based (w, g) = Σs∈D |w|s |g − w|s, (3)

Neighbors (g) =Σs∈D Σ w∈s |w|s |g − w|s. (4)

|g − w|s = {|g|s − |w|s, w ∈ g, (5)
{|g|s, w!∈g

Extracting roofs: The strength of a column between high key term wi and high-frequency term wj ⊂ HF is expressed as

Column (wi, wj) = Σs∈D min (|wi|s, |wj |s). (6)

Columns touching wi are sorted by column (wi, wj) for each high key term wi. Columns with the highest column values connecting term wi to two or more clusters are selected to create new links in G. Finally, the nodes in G are sorted by the sum of the column of the touched columns. Terms represented by the nodes of higher values than a certain threshold are extracted as the keywords for document D.

3. RESULTS AND ANALYSIS

Table 1: Relationship between Top Ten Words and Their Frequency for Internal Medicine.

Rank	Internal Medicine 1	Frequency	Internal Medicine 2	Frequency	Internal Medicine 3	Frequency
1	Doctor	172	Internal use	136	Description	13
2	Direction	140	Medication	84	Charge	13
3	Principal	139	Connection	72	Alcer	13
4	patient	139	Urine	70	Acne	13
5	Internal use	131	Feeling	60	Skin	12
6	Report	130	Today	59	Toady	11
7	Hope	102	Principle	54	Use	10
8	Consent	92	Attending Physician	54	Entire body	9
9	Observation	91	Progress	91	Plan	9
10	Use	89	Hope	51	Exchange	9

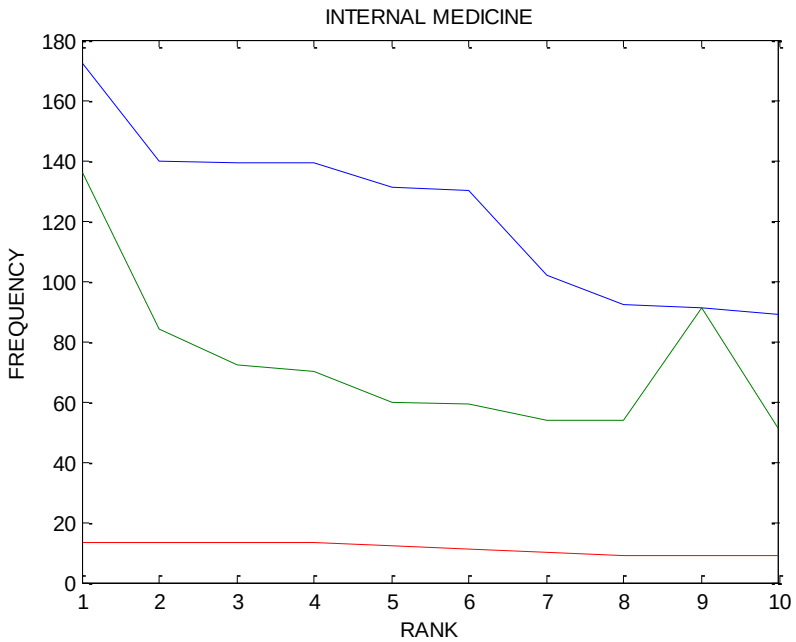


Fig 4: Internal Medicine.

Table 2: Relationship between Top Ten Words and Their Frequency for Surgery and Ophthalmology.

Rank	Surgery 1	Frequency	Surgery 2	Frequency	Ophthalmology	Frequency
1	Acne	54	Internal use	514	Eye	71
2	Ulcer	49	Principal	265	Internal use	64
3	Change	49	Sleep	245	Complaint	55
4	Skin	48	Use	243	Trial	47
5	Reddening	44	Complaint	200	Terminology	43
6	Description	43	Description	195	Today	40
7	Entire body	35	Hope	186	Direction	40
8	Plan	35	Progress	183	Rest	38
9	Necessary	34	Urine	170	Sore	37
10	Continuation	32	Then	169	Surgery	36

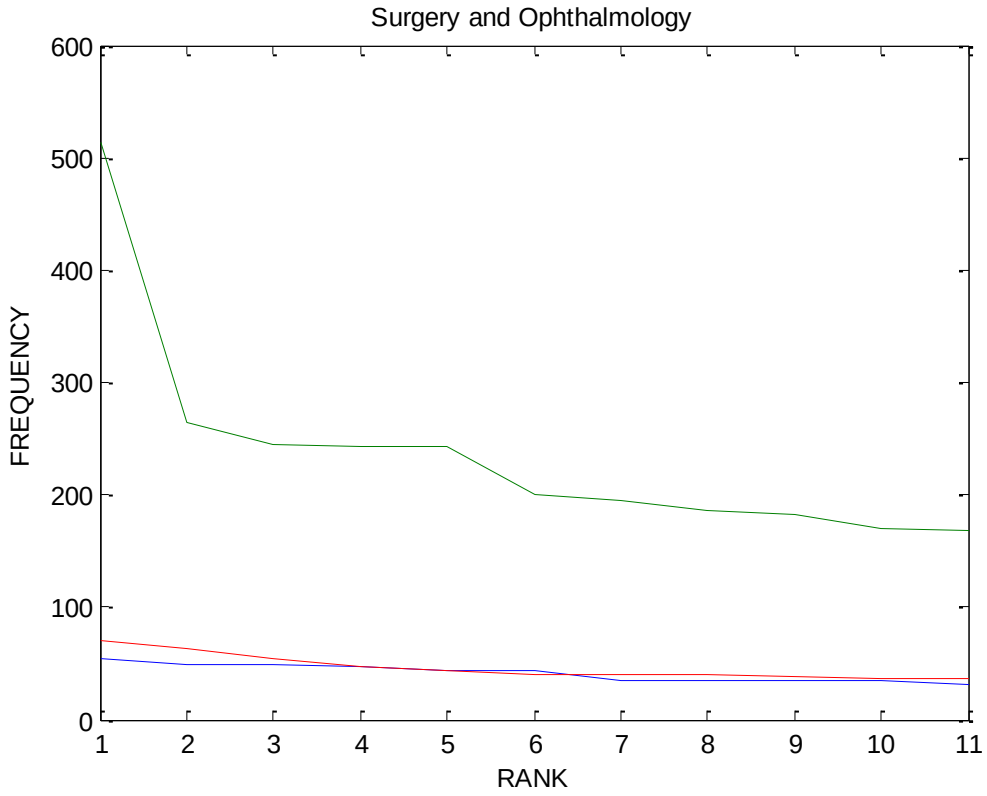


Fig 5: Surgery and Ophthalmology.

Table 3: Relationship between Top Ten Words and Their Frequency for Obstetrics and Gynecology, Oral and Maxillofacial Surgery and Otorhinolaryngology.

Rank	Obstetrics and Gynecology	Frequency	Oral and Maxillofacial Surgery	Frequency	Otorhinolaryngology	Frequency
1	Sore	99	Admission	99	Pressure	102
2	Internal use	82	Blood vessel	98	Shade	101
3	Complaint	80	Punctures	98	Destruction	98

4	Direction	74	Connection	98	Right	88
5	Today	68	Leakage	97	Neck	85
6	Bleeding	68	Nursing	97	Left	72
7	Doctor	68	Reference	92	Blood	68
8	Report	62	Infection	74	Chest	45
9	Above	61	Plan	71	Progress	41
10	Uterus	56	Attachment	71	Reference	40

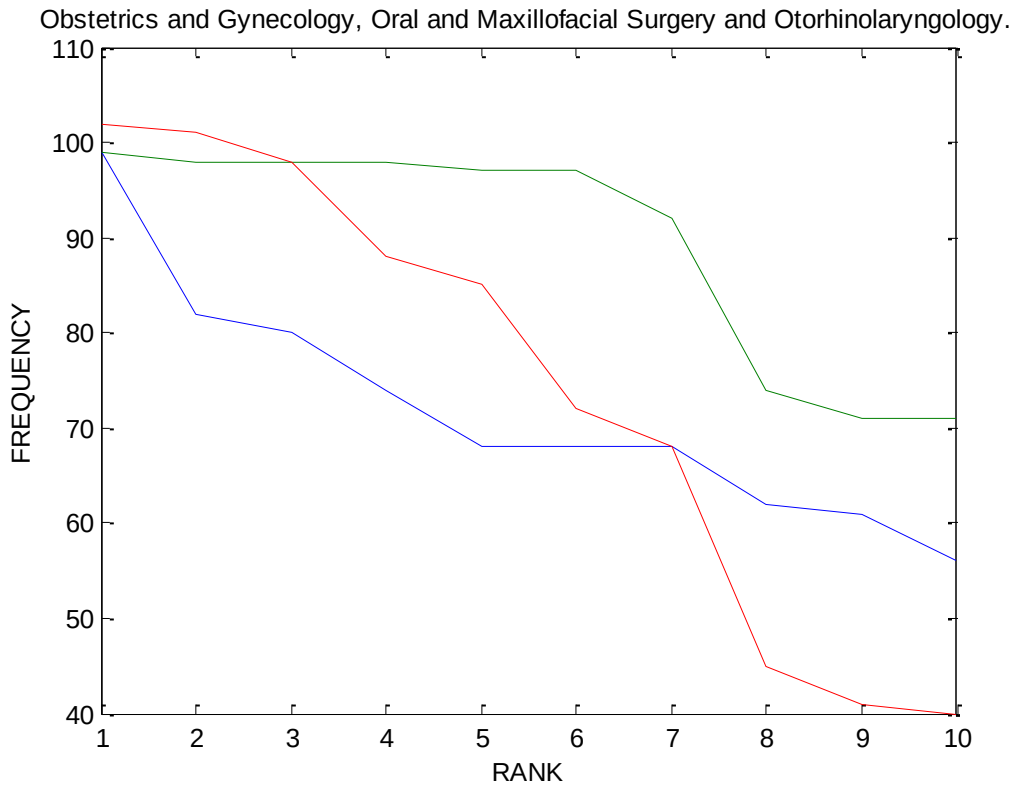


Fig 6: Obstetrics and Gynecology, Oral and Maxillofacial Surgery and Otorhinolaryngology.

Table 4: Relationship between Top Ten Words and Their Frequency for Pediatrics, Orthopedics, and Psychiatry.

Rank	Pediatrics	Frequency	Orthopedics	Frequency	Psychiatry	Frequency
1	Medical treatment	23	Internal use	431	Internal use	239
2	Description	22	Sleep	186	State	208
3	Nursing	20	Hope	119	Sleep	193
4	Doctor	20	Principal	117	Arousal	173
5	Suffer	19	Medicine	110	Then	167
6	State	19	Complaint	99	Check	156
7	Necessary	17	Above mentioned	71	Inspection tour	136
8	Mother	17	Use	74	Night	107
9	Breathing	17	Trial	67	Complaint	106
10	Family	16	Description	65	Principal	97

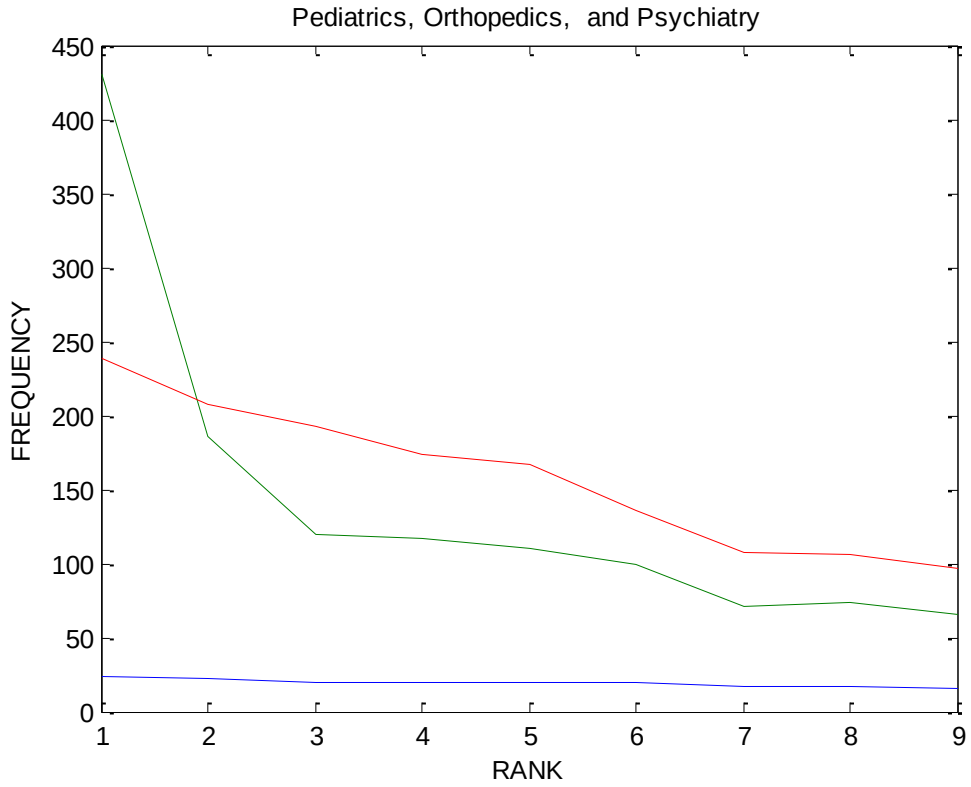


Fig. 7: Pediatrics, Orthopedics, and Psychiatry.

Table 5: Relationship between Top Ten Words and Their Frequency for Neurosurgery, Urolog and Dermatology.

Rank	Neurosurgery	Frequency	Urology	Frequency	Dermatology	Frequency
1	Internal use	137	Internal use	34	Internal use	130
2	Then	115	Principal	24	Pain	76
3	Complaint	109	Bloody urine	20	Hope	54
4	Direction	95	Complaint	20	Progress	49
5	Consciousness	93	Urine	19	Refererences	45
6	Doctor	84	Use	18	Principal	36
7	Sore	83	Description	18	Record	36
8	Progress	83	Hope	18	More	35
9	Description	77	Skin	17	Patient	34
10	Above mentioned	75	Operation	17	Sore	34

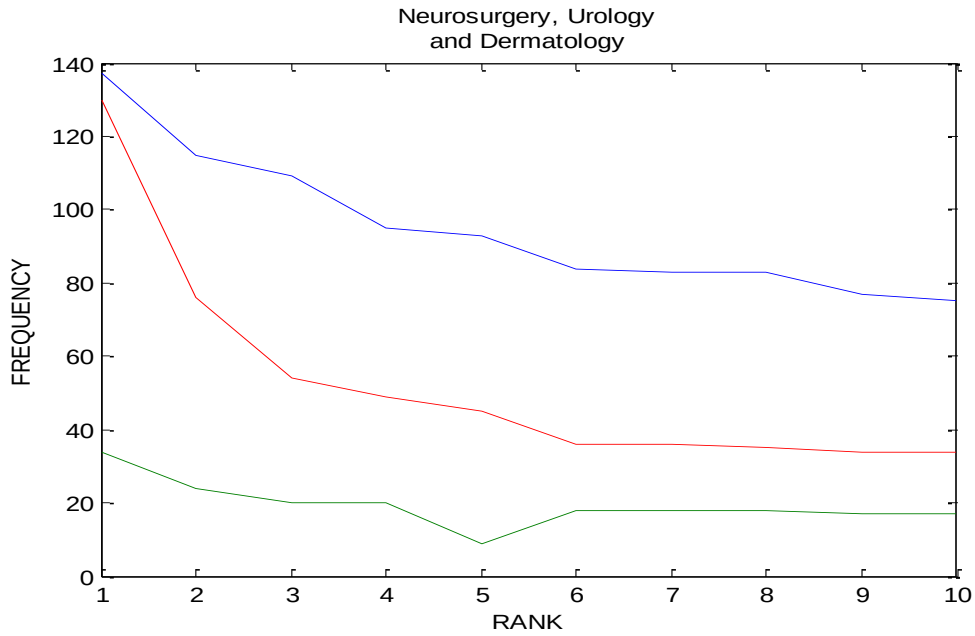


Fig. 8: Neurosurgery, Urology and Dermatology.

Table 6: Relationship between Top Ten Words and Their Frequency for Radiology, Anesthesiology and Rheumatology and Infections.

Rank	Radiology	Frequency	Anesthesiology	Frequency	Rheumatology and Infectious Diseases	Frequency
1	Blood vessel	147	Infusion	116	Internal use	68
2	Leakage	46	Punctures	115	Urine	65
3	Admission	46	Admission	105	Progress	61
4	Punctures	46	Rest	113	Breathing	52
5	Connection	46	Connection	111	Oxygen	49
6	Tip	45	Fixation	110	All right	37
7	Rest	44	Pollution	110	Use	37
8	Lft hand	43	Nursing	32	Inhalation	35
9	Nursing	38	Content	28	Night	34
10	Plan	29	Attendance	27	Principal	33

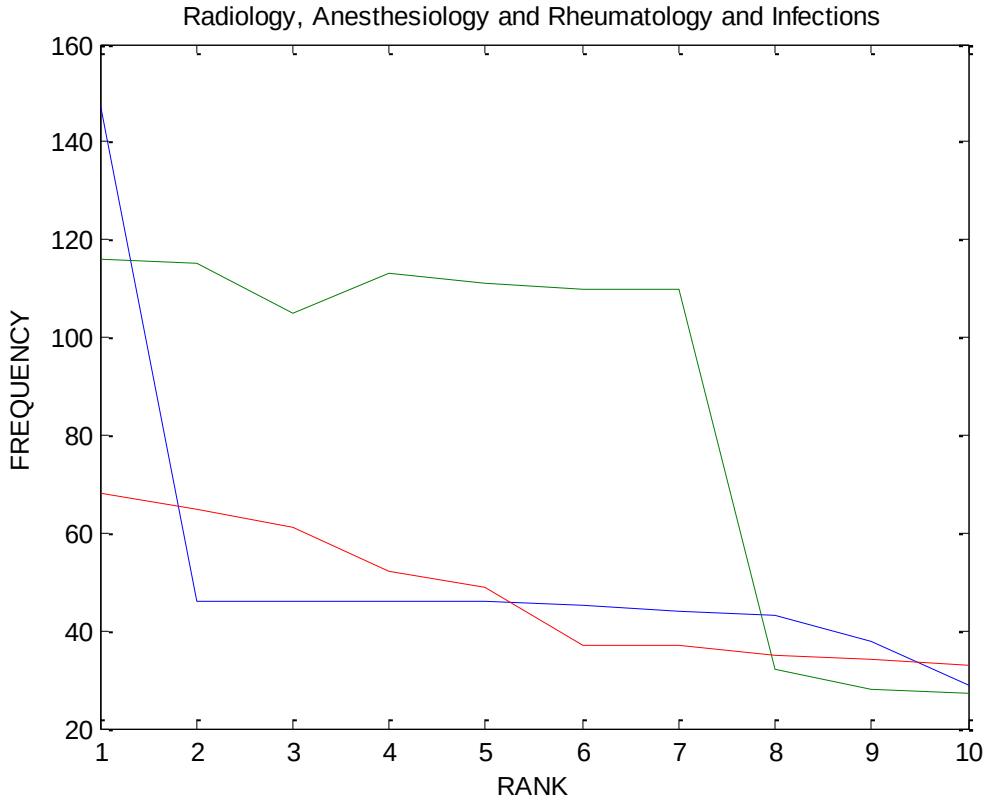


Fig. 9: Radiology, Anesthesiology and Rheumatology and Infections.

4. CONCLUSIONS

The selection of right and appropriate text mining technique helps to enhance the speed and decreases the time and effort required to extract valuable information. The analysis results have potential to clarify the work content of care workers. As the nursing field requires efficiency in health care services, improvement and continuous data collection are important for the long-term building of health care services as well as large-scale data collection**.**

5. Scope of Future work

In the future, we aim to develop an Electronic Medical Record that can be created semi-automatically in accordance with the level of care required.

REFERENCES

1. A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravi˘cius, and M. Duneld, “Synonym extraction and abbreviation expansion with ensembles of semantic spaces,” Journal of biomedical semantics, vol. 5, International Conference on. IEEE, 2013; 1(1): 78–81.

2. A. Henriksson, J. Zhao, H. Dalianis, and H. Bostr¨om, “Ensembles of randomized trees using diverse distributed representations of clinical events,” BMC Medical Informatics and Decision Making, 2016; 16(2): 69.
3. A. Kaklauskas, M. Seniut, D. Amaratunga, I. Lill, A. Safonov, N. Vatin, J. Cerkauskas, I. Jackute, A. Kuzminske, and L. Peciure, “Text analytics for android project,” Procedia Economics and Finance, 2014; 18: 610–617.
4. A. Kumaran, R. Makin, V. Pattisapu, and S. E. Sharif, “Automatic extraction of synonymy information:-extended abstract,” OTT06, 2007; 1: 55.
5. A. M. Cohen and W. R. Hersh, “A survey of current work in biomedical text mining,” Briefings in bioinformatics, 2005; 6(1): 57–71.
6. B. A. Mukhedkar, D. Sakhare, and R. Kumar, “Pragmatic analysis based document summarization,” International Journal of Computer Science and Information Security, 2016; 14(4): 145.
7. B. L. Narayana and S. P. Kumar, “A new clustering technique on text in sentence for text mining,” IJSEAT, 2015; 3(3): 69–71.
8. B. Laxman and D. Sujatha, “Improved method for pattern discovery in text mining,” International Journal of Research in Engineering and Technology, 2013; 2(1): 2321–2328.
9. C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” Journal of bioinformatics and computational biology, 2005; 3(02): 185–205.
10. C. P. Chen and C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on big data,” Information Sciences, 2014; 275: 314–347.
11. D. S. Dang and P. H. Ahmad, “A review of text mining techniques associated with various application areas,” International Journal of Science and Research (IJSR), 2015; 4(2): 2461–2466.
12. E. A. Calvillo, A. Padilla, J. Mu˜noz, J. Ponce, and J. T. Fernandez, “Searching research papers using clustering and text mining,” in Electronics, Communications and Computing (CONIELECOMP), 2013.
13. F. Fatima, Z. W. Islam, F. Zafar, and S. Ayesha, “Impact and usage of internet in education in pakistan,” European Journal of Scientific Research, 2010; 47(2): 256–264.
14. G. King, P. Lam, and M. Roberts, “Computer-assisted keyword and document set discovery from unstructured text,” Copy at http://j. mp/1qdVqhx Download Citation BibTex Tagged XML Download Paper, 2014; 456.

15. H. Solanki, "Comparative study of data mining tools and analysis with unified data mining theory," International Journal of Computer Applications, 2013; 75(16).
16. I. Alonso and D. Contreras, "Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An umls approach," Expert Systems with Applications, 2016; 44: 386–399.
17. I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan, "Text mining in a digital library," International Journal on Digital Libraries, 2004; 4(1): 56–59.
18. K. Sumathy and M. Chidambaram, "Text mining: Concepts, applications tools and issues-an overview," International Journal of Computer Applications, 2013; 80(4).
19. N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in Proceedings of the World Congress on Engineering, 2013; 3: 3–5.
20. N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," IEEE transactions on knowledge and data engineering, 2012; 24(1): 30–44.
21. P. J. Joby and J. Korra, "Accessing accurate documents by mining auxiliary document information," in Advances in Computing and Communication Engineering (ICACCE), Second International Conference on. IEEE, 2015; 634–638.
22. R. Agrawal and M. Batra, "A detailed study on text mining techniques," International Journal of Soft Computing and Engineering (IJSCE) ISSN, 2013; 2231–2307.
23. R. Al-Hashemi, "Text summarization extraction system (tses) using extracted keywords." Int. Arab J. e-Technol., 2010; 1(4): 164– 168.
24. R. Rajendra and V. Saransh, "A Novel Modified Apriori Approach for Web Document Clustering," International Journal of Computer Applications, 2013; 159–171.
25. R. Sharda and M. Henry, "Information extraction from interviews to obtain tacit knowledge: A text mining application," AMCIS Proceedings, 2009; 283.
26. R. Steinberger, "A survey of methods to ease the development of highly multilingual text mining applications," Language Resources and Evaluation, 2012; 46(2): 155–176.
27. S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan, "Data mining model for higher education system," Europen Journal of Scientific Research, 2010; 43(1): 24–29.
28. S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques and applications–a decade review from 2000 to 2011," Expert Systems with Applications, 2012; 39(12): 303–11-311.
29. V. Gupta and G. S. Lehal, "A survey of text mining techniques and applications," Journal of emerging technologies in web intelligence, 2009; 1(1): 60–76.

30. W. He, “Examining students online interaction in a live video streaming environment using data mining and text mining,” Computers in Human Behavior, 2013; 29(1): 90–102.
31. Y. Zhao, “Analysing twitter data with text mining and social network analysis,” in Proceedings of the 11th Australasian Data Mining and Analytics Conference (AusDM 2013), 2013; 23.
32. Z. Wen, T. Yoshida, and X. Tang, “A study with multi-word feature with text classification,” in Proceedings of the 51st Annual Meeting of the ISSS-2007, Tokyo, Japan, 2007; 51: 45.