Diversifications of Topics in Journal Articles: A Data Driven Analysis from 2000 - 2022
Text Mining Course Work
Group: 1(a)

by

34624309
34749497
35496215

School of Social Sciences
Faculty of Mathematical Sciences
University of Southampton
12 January 2024

Abstract

This study employs text mining identify top words, search for varied topics or text clusters and predict citations and types of articles among different journals. Using methods like Bag of Words, TF-IDF, LDA for topic modeling, and a regression model, we analyzed a dataset of journal abstracts for thematic trends and predictive features. The results validate our hypothesis, demonstrating that specific textual elements correlate with citation counts. This analysis provides insights into evolving research trends and highlights text mining's role in bibliometric studies, offering implications for researchers and publishers in enhancing academic impact.

Keywords: text predicts citations, journals, words reveal focus, algorithm dictate themes, data driven analysis, bag of words, TF-IDF, topic modelling, regression, classification, association, clustering, PCA

0.1 Highlight

1. Among different clustering methods, we found that Clara is the best algorithm to cluster the dataset, which is summarized after many failures.

2. The Journal of Simulation covers a wider range of themes, while Health Systems and the Journal of the Operational Research Society focus further analysis within specific topics.

3. We implemented multiple classification to help identify different types of journals and over 90% of articles in the testing set can be correctly classified.

4. The words "health", "care", "patient", "disease", "surgery", "agent", "validation", "conceptual", "behavior", "electronic" are the most important ten words to distinguish different journal types.

5. SVM model accurately predicts article citations with a clear linear fit, affirming text mining efficacy.

Table of Contents

Chapter 1

Introduction

1.1 Introduction

In our report, we delve into text analysis with a dataset of 4,385 journal articles from Health System, Journal of Simulation, and Journal of the Operational Research Society, spanning 2000-2022. This rich dataset allows us to examine research trends and methodologies over two decades. Our goal is to analyze the textual content of these articles, exploring the relationship between the language used in abstracts and their citation impact. This study demonstrates the utility of text mining in identifying patterns and tracking the evolution of academic research.

1.2 Hypothesis

Our hypothesis posits that a classification model can accurately predict article types from abstracts, document clustering is related to journal types, and current citations can be predictive of future citations.

Therefore, we want to test our hypothesis by attempting to answer the following questions in our work:

Is it possible to create a classification model that predicts the type of a article and assigns it to the appropriate journal based on its abstract? Does document clustering have certain relationship with journal type? And can we predict citations based on our current citation?

Chapter 2

Data Collection and Preprocessing

2.1 Introduction to Data Collection

This chapter describes data collection and preprocessing for analyzing 4384 abstracts from Health System, Journal of Simulation, and Journal of the Operational Research Society, published between 2000 and 2022.

2.2 Data Preprocessing and Analysis

In the Data Preprocessing and Analysis phase, we standardized text data, eliminated stopwords, and applied stemming and lemmatization to simplify words to their roots. Subsequently, we produced word clouds for each journal, visually emphasizing key words and themes in the abstracts, integral for our analysis.

2.3 Word Clouds

Word clouds were created to visualize the most frequently occurring words within the abstracts of each journal.

These visual representations provide an at-a-glance understanding of the data's textual content:

1. Health System Journal Abstracts:
   The word cloud for the Health System journal highlights a strong prevalence of terms directly related to medical practice and patient care. Words such as "patient," "health," "care," "hospital," "healthcare," and "service" are prominently displayed, indicating a clear focus on healthcare delivery and patient-centric topics within this journal.

Fig.3.3.1: Health System

2. Journal of Simulation Abstracts:
   For the Journal of Simulation, the word cloud reveals a lexicon centered around modeling and computational techniques. Key terms include "network," "agent," "discrete," "dynamic," "application," and "algorithm," signifying the journal's emphasis on simulation methodologies and their applications in various domains.

Fig.3.3.2: Journal Of Simulation

3. Journal of the Operational Research Society Abstracts:
   The abstracts from the Journal of the Operational Research Society frequently contain words such as "algorithm," "solution," "cost," "optimal," "strategy," and "measurement." These terms suggest a strong orientation towards problem-solving, cost-efficiency, and strategic planning within the research covered by this journal.

Fig.3.3.3: Journal Of Operational Research

Chapter 3

Analysis of Top Words' Trend

3.1 Analysis of Top Words' Trend

After creating word clouds, we quantified the top words' frequency over time and across journals, validating the word clouds visually and offering measurable insights into their prevalence and significance.

3.1.1 Trending Words of Each Year in Different Journals

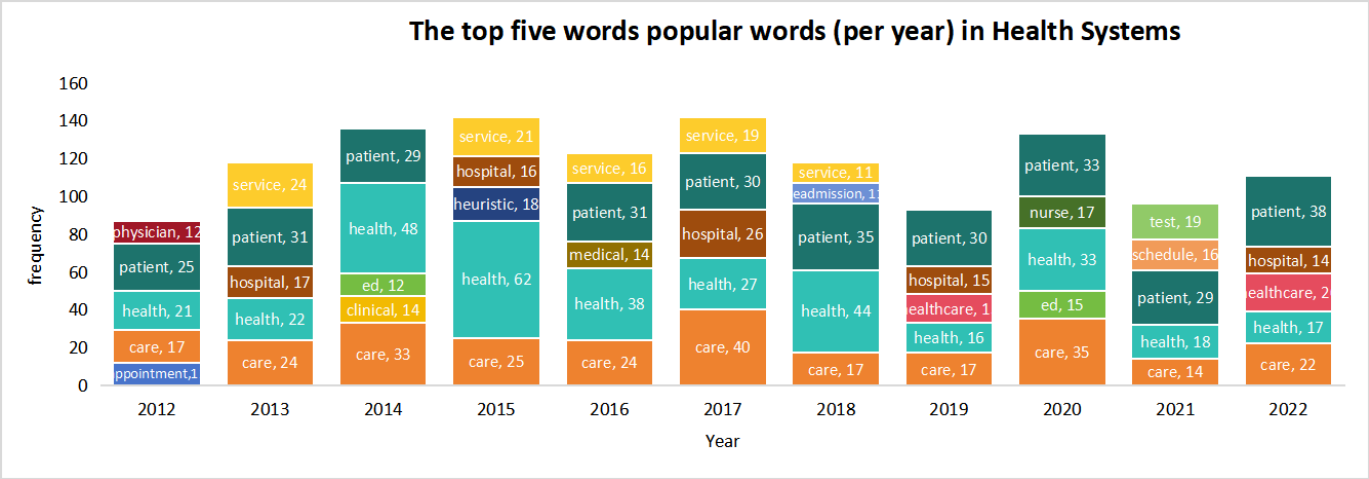

Fig.3.1.1: The top five popular words(per year) in Health System

The bar chart shows the top five words yearly in Health Systems journals from 2012 to 2022, highlighting "patient," "health," "care," and "service" as dominant. These terms underline the journal's emphasis on patient-focused care and healthcare services. Their consistent use over the years mirrors evolving research interests in healthcare, with a noticeable focus on service optimization in 2012 and continued attention to patient care and health management.

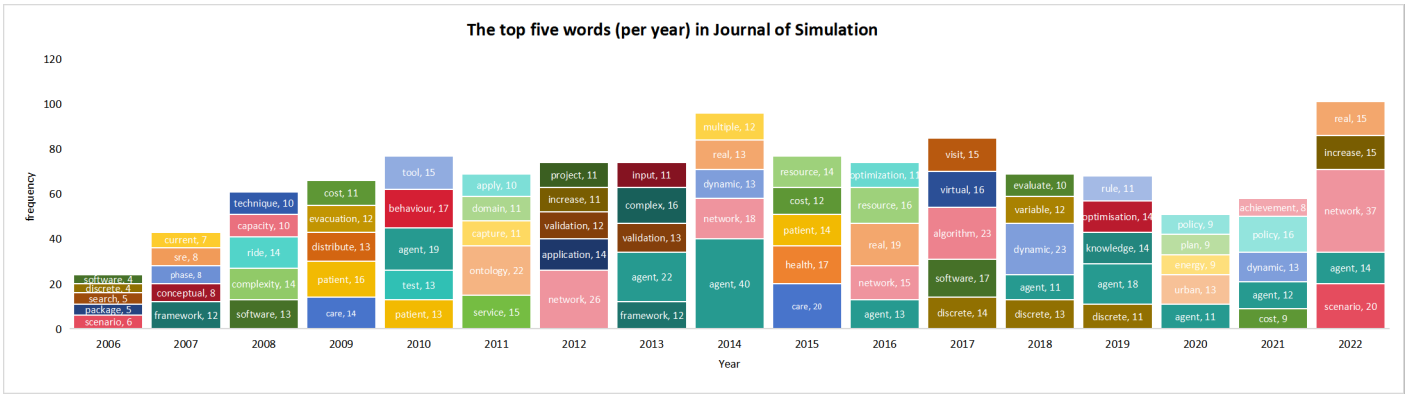

Fig.3.1.2: The top five popular words(per year) in Journal Simulation

The Journal of Simulation's dataset from 2006 to 2022 shows varied topics, with frequent terms like "network," "agent," and "policy." In 2015, a focus on health emerged, with 8 out of 29 articles featuring health themes, reflecting an interest in applying simulation to healthcare. By 2020, the focus shifted to "energy" and "urban" themes, mirroring global trends in energy and city planning, showcasing the journal's adaptability to evolving real-world issues.

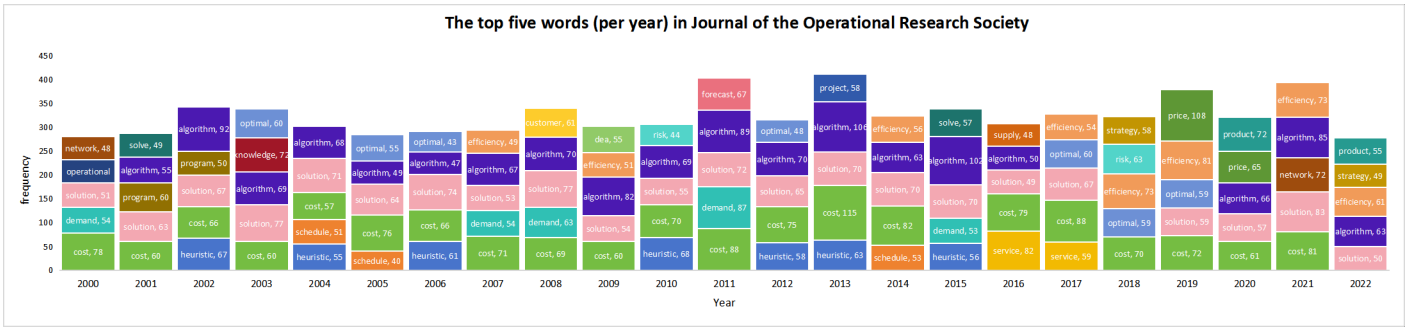

Fig.3.1.3: The top five popular words(per year) in Operational Research Society

From 2000 to 2022, the Journal of the Operational Research Society consistently used terms like "algorithms," "solutions," and "cost," emphasizing problem-solving and economic aspects of operational research. Since 2017, "efficiency" became prominent, indicating a shift towards optimizing resources and system performance.

Comparatively, the Journal of Simulation covers a wider range of themes, while Health Systems and the Journal of the Operational Research Society focus more narrowly, suggesting deeper exploration within specific topics. This difference in thematic focus may influence the context and frequency of their top words.

Chapter 4

Text Mining Models

4.1 Topic Modelling

Using LDA, we identified distinct topics across journals: Topic 1, with "supply," "chain," and "demand," pertains to operational research and logistics, while Topic 3, featuring "system," "health," "patient," and "care," focuses on healthcare systems and patient care, revealing the thematic diversity in these journals.

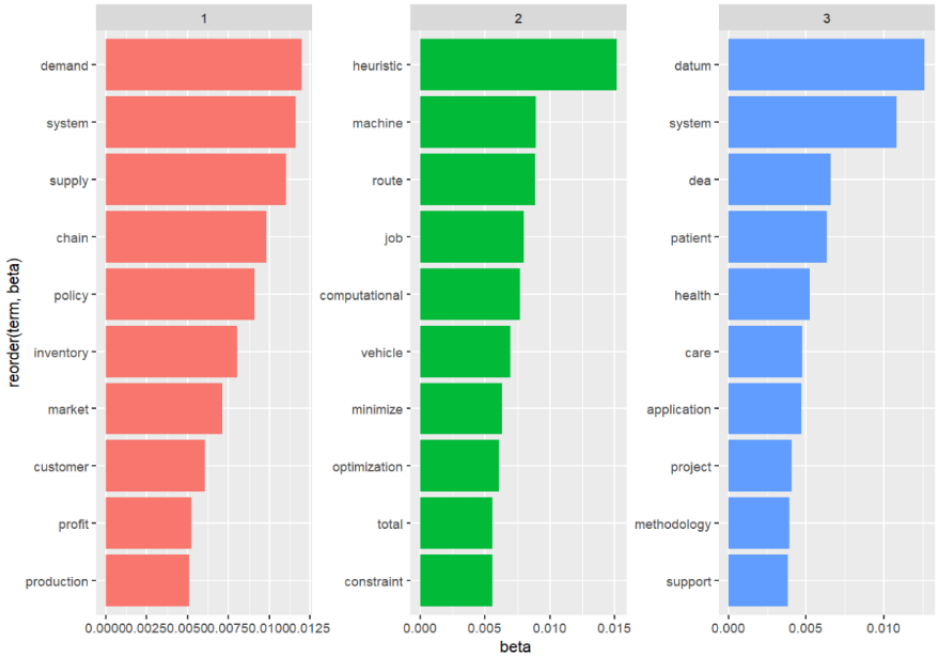

Fig.4.1: Top Words

4.2 Significance of Words to Topics

To gauge word importance in LDA topics, we computed log ratios of β values for topics 1, 2, and 3, targeting words with β >0.01 and log ratios >30. The plot shows high positive ratios indicating significance in Topic 1, and negative ratios in Topics 2 or 3. Key words like "flowshop" and "dmus" define Topic 1 (operational research), while "patient" and "care" dominate Topic 3 (health systems), aiding in thematic

understanding.

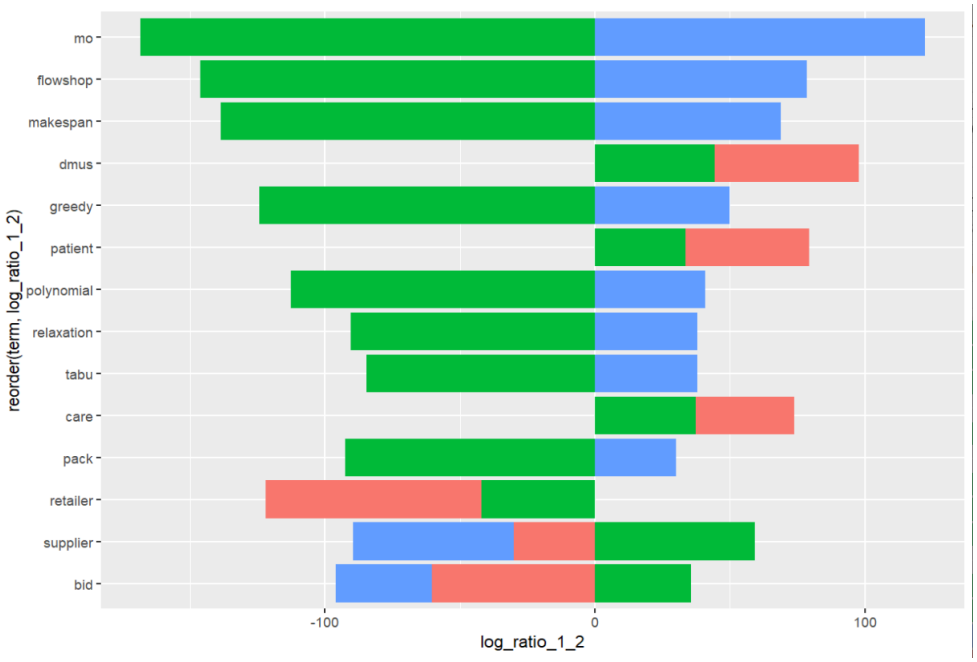

Fig.4.2: Log Ratio

4.3 Distribution of Topics Across Documents

We set a γ textgreater 0.95 to find each document's main topic. The bar chart shows Topic 3's dominance, featuring "patient," "health," and "care," while Topic 1's "supply" and "demand" occur less, indicating its niche focus. This suggests that broader topics like healthcare are more widely discussed in academic discourse.

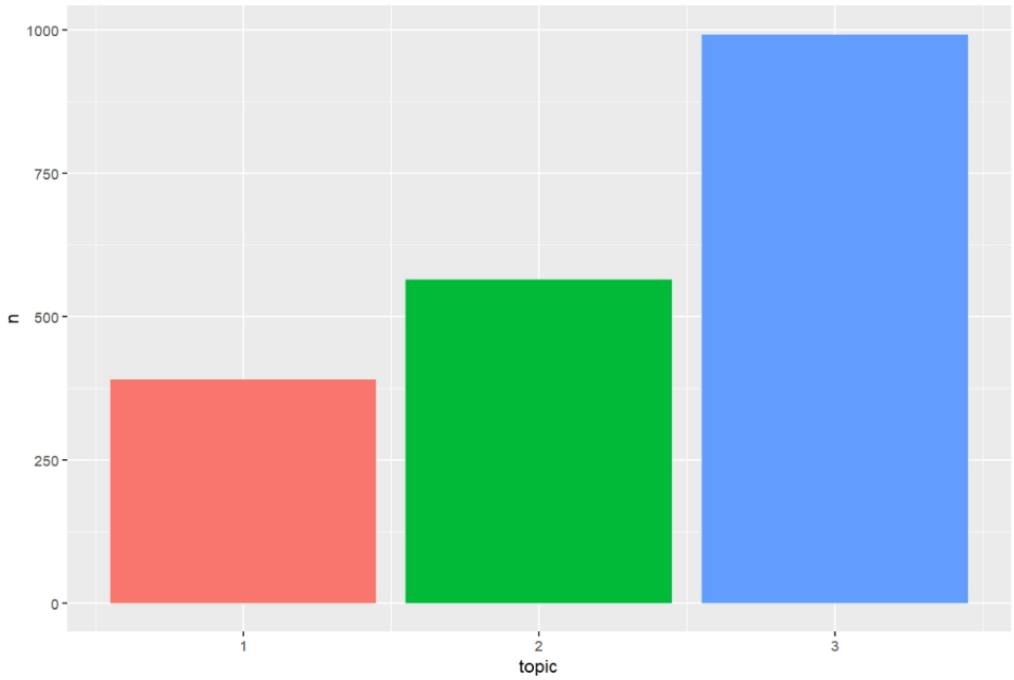

Fig.4.3: Topics Distribution

Chapter 5

Clustering and PCA

5.1 Clustering Techniques Overview

To uncover inherent structures in the dataset, we explored various clustering methods. Initial trials with k-means and PAM showed complex, disordered distributions, indicating these methods' ineffectiveness for our large, high-dimensional dataset.

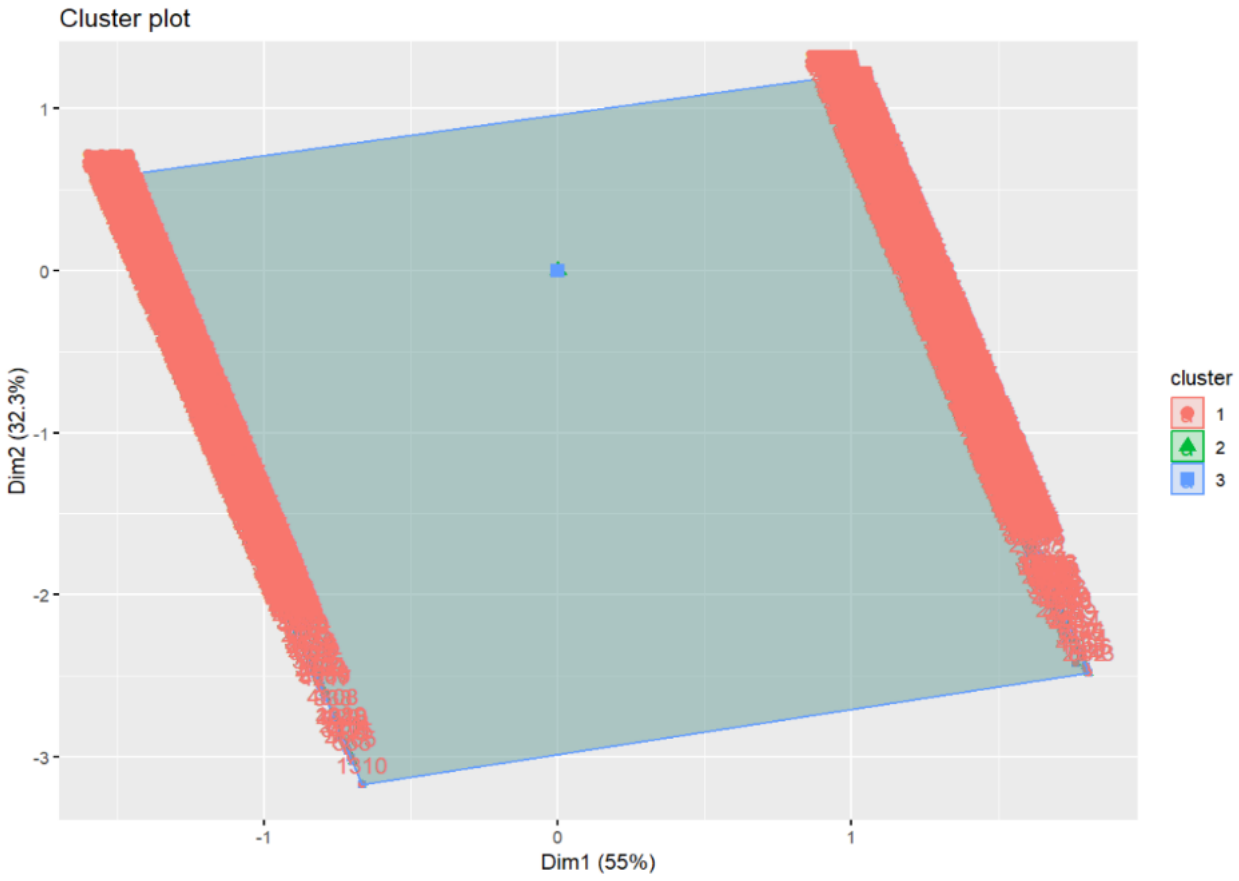

Fig.5.1.1: Cluster Plot using K-means

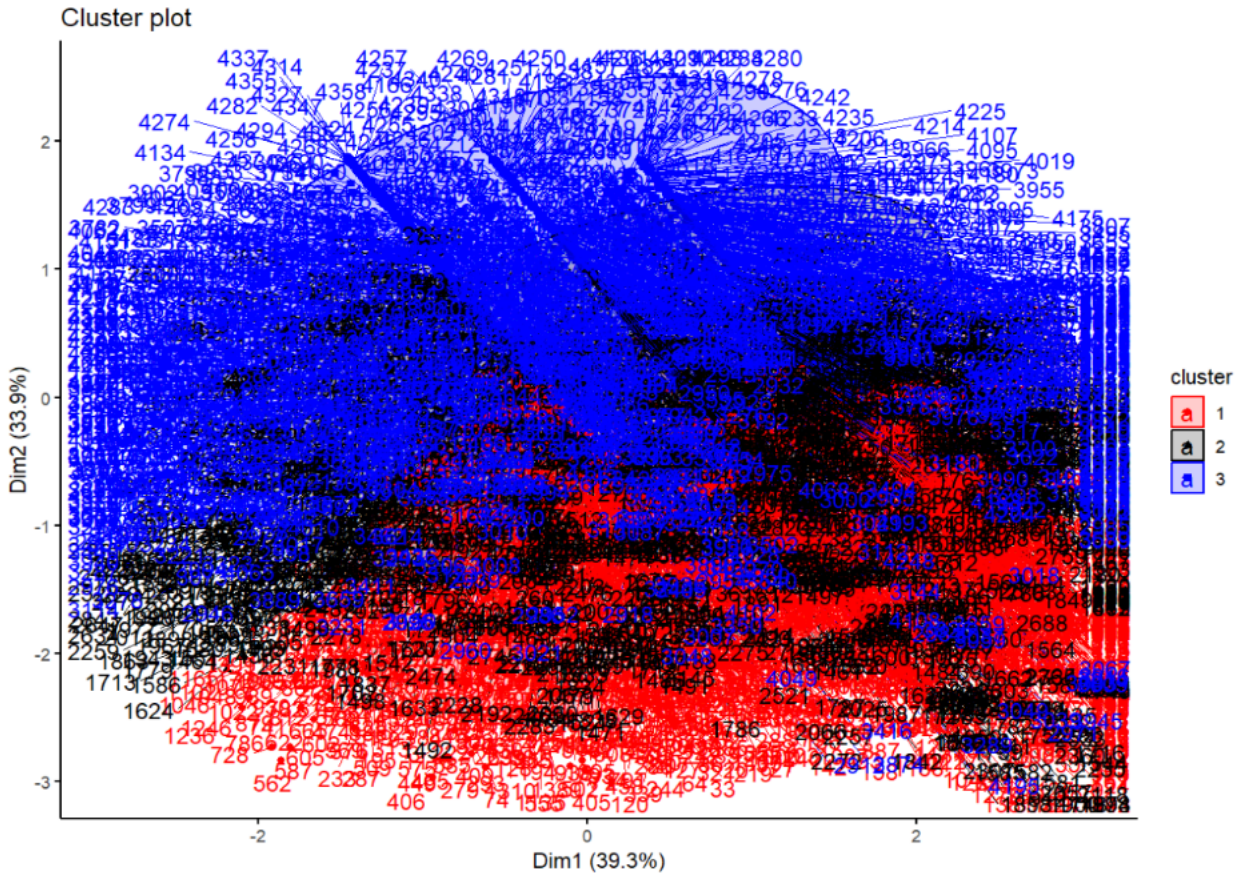

Fig.5.1.2: Cluster Plot using PAM

5.2 Selection of CLARA Algorithm

Facing challenges with initial methods, we chose the CLARA (Clustering Large Applications) algorithm, suited for large datasets via sampling. CLARA yielded clearer cluster separation, distinguishing three distinct groups in the dataset, as shown in the third cluster plot.

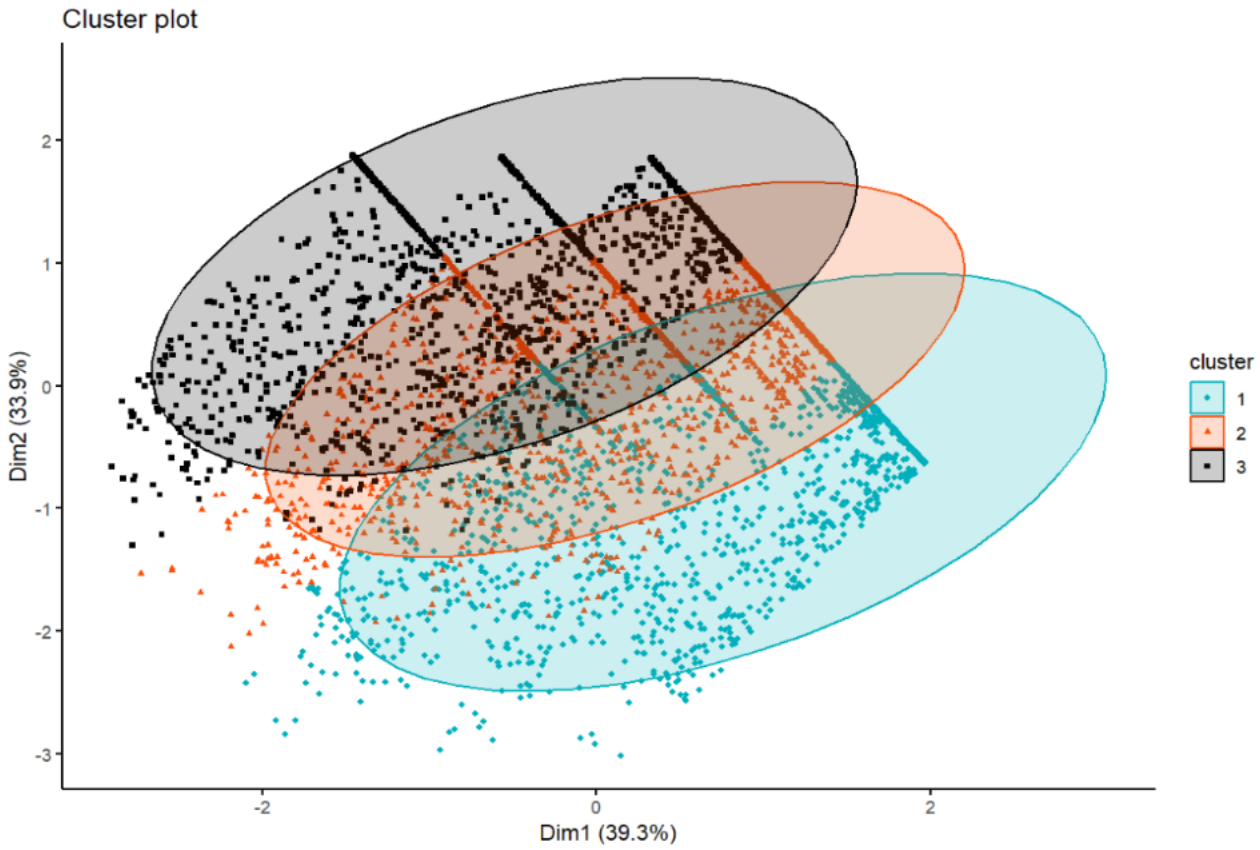

Fig.5.2.1: Cluster Plot using Clara

5.3 Visualization with PCA

PCA (Principal Component Analysis) was used to reduce the dataset to two principal components for easier interpretation and 2D visualization, making the clusters distinctly visible and simplifying the understanding of the data's structure.

5.4 Analysis of Cluster Sizes

Further analysis on cluster size and composition revealed a balanced distribution, with each cluster containing about 1,500 documents, indicating no inherent bias towards any specific cluster.

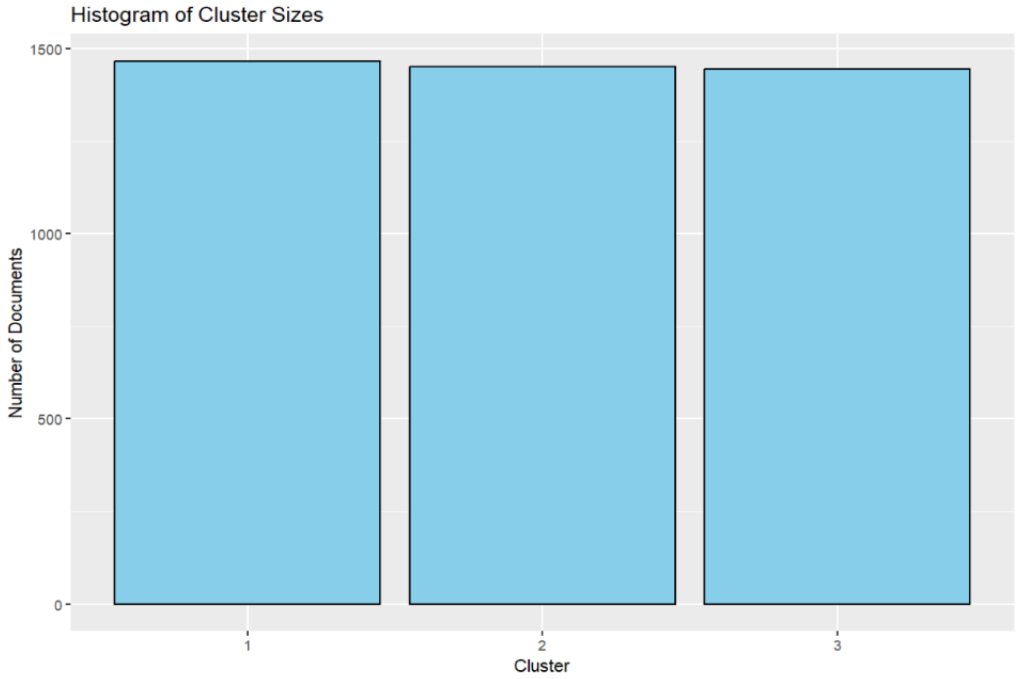

Fig.5.4.1: Histogram of Number of Documents using Clara

5.5 Topics Distribution Among Clusters

In the final step, we assigned topics to each cluster by matching them with topic probabilities from the LDA model. This identified each cluster's most representative topic, allowing us to label clusters based on their prevalent words and dominant themes.

label	topic1	topic2	topic3	sum
cluster1	305	358	803	1466
cluster2	335	618	498	1451
cluster3	611	413	420	1444
sum	1251	1389	1721	4361

Fig.5.5: Histogram of Number of Documents using Clara

Fig.5.6: Topic Distribution Bar Chart comparison between Clusters

Chapter 6

Association and Correlation

6.1 Pairwise Associations

In the articles of the Health System, “patient”, “care”, and “health” appeared in pairs more than 200 times in all abstracts. As for the rest types of journal, the words “application”, “discrete” and “tool” appeared in pairs about 50 times in the abstracts of the Journal of Simulation. For the words “solution” in the papers of the Operational Research Society, it co-occurs more than 1,600 times in pairs with different words, which include “algorithm”, “optimal”, “heuristic” and so on.

Health Systems			Journal of Simulation			Journal of the Operational Research Society		
Words1	Words2	n	Words1	Words2	n	Words1	Words2	n
care	health	79	application	discrete	25	algorithm	solution	374
health	care	79	discrete	application	25	solution	algorithm	374
care	patient	66	real	world	22	solve	solution	335
patient	care	66	discrete	tool	22	solution	solve	335
health	patient	55	tool	discrete	22	algorithm	solve	320
patient	health	55	world	real	22	solve	algorithm	320
hospital	patient	46	dynamic	behaviour	21	solution	optimal	293
patient	hospital	46	behaviour	dynamic	21	optimal	solution	293
hospital	care	40	support	tool	20	heuristic	solution	289
service	care	40	dynamic	agent	20	solution	heuristic	289
...	...	...	...	...	...	...	...	...
498,388 pairs of words			716,238 pairs of words			3,440,972 pairs of words		

Fig.6.1.1: Relation of Special Words

6.2 Extensive Word Correlations Across Journals

Apart from the pairwise counting of the top 10 terms as previously noted, more correlations between top words were explored by plotting IMDB graphs, which are expected to show us more about how more than a special word is related to many other words.

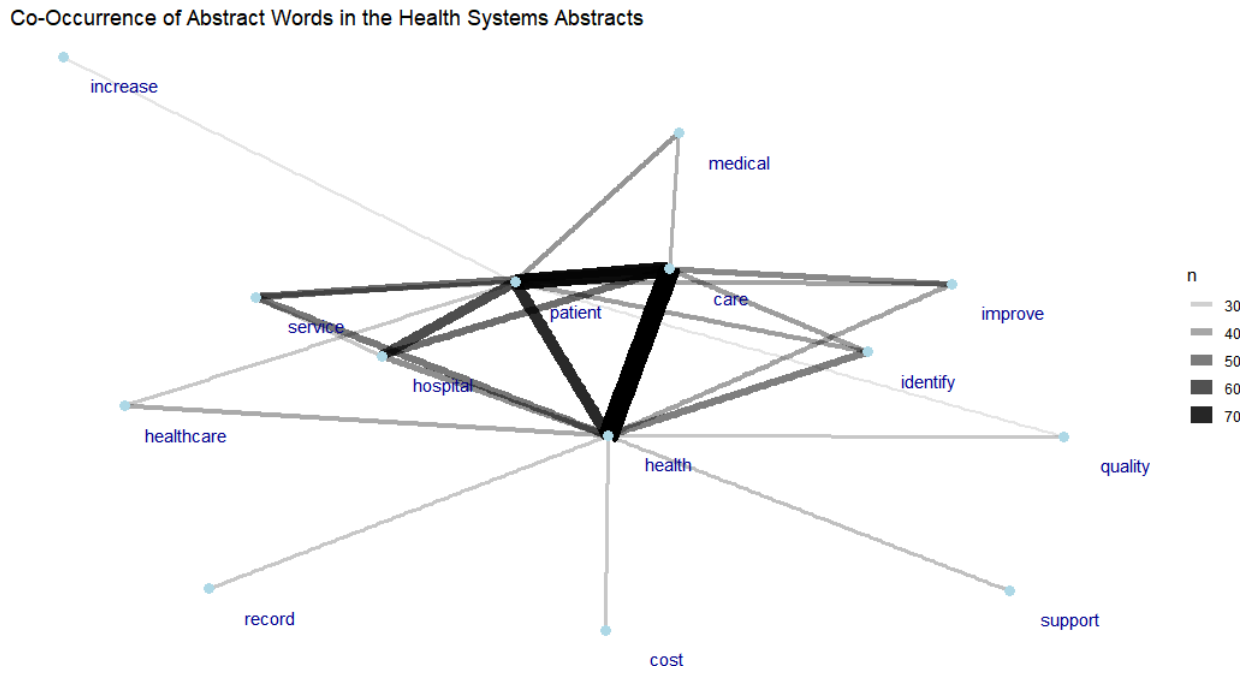

Fig.6.2.2: Co-Occurence Health System

The Health Systems co-occurrence graph links "patient" and "care" with "medical," "improve," and "identify," showing a focus on medical care and patient identification, underscoring a drive for quality care and understanding patient needs, and confirming a dedication to medical advancement.

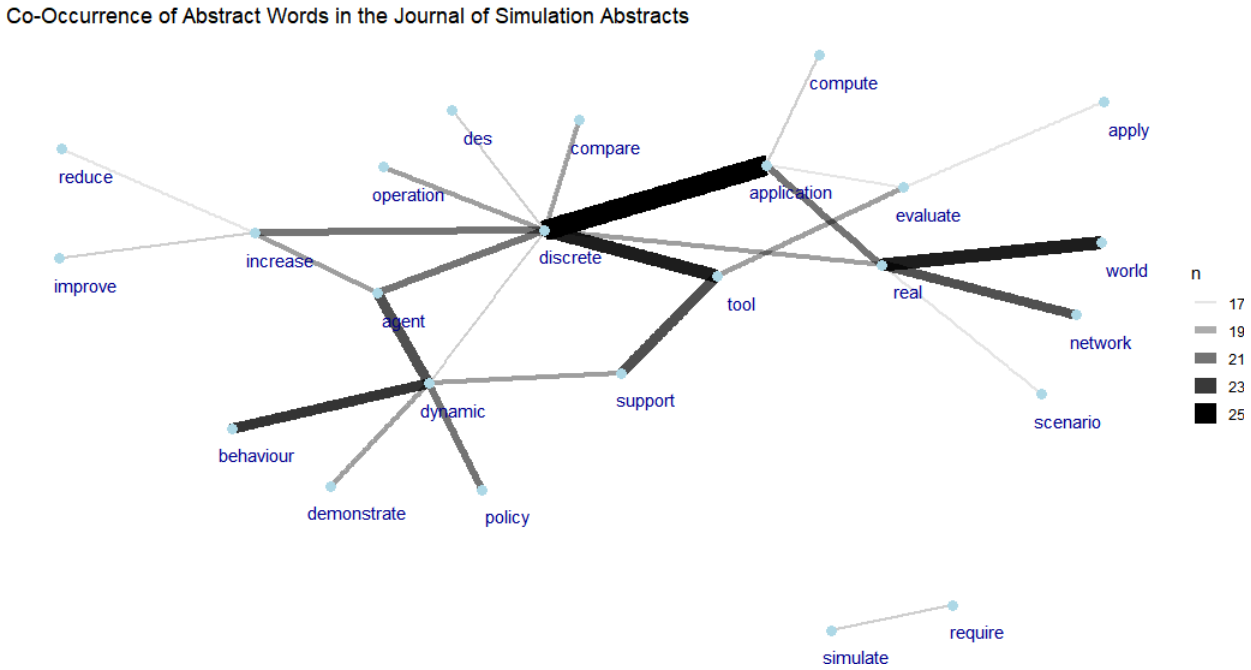

Fig.6.2.3: Co-Occurrence Journal System

In the Journal of Simulation, "discrete" associates with "application," "tool," "compare," "operation," and "agent," emphasizing its focus on discrete event simulation and agent-based modeling for operational efficiency in operational research.

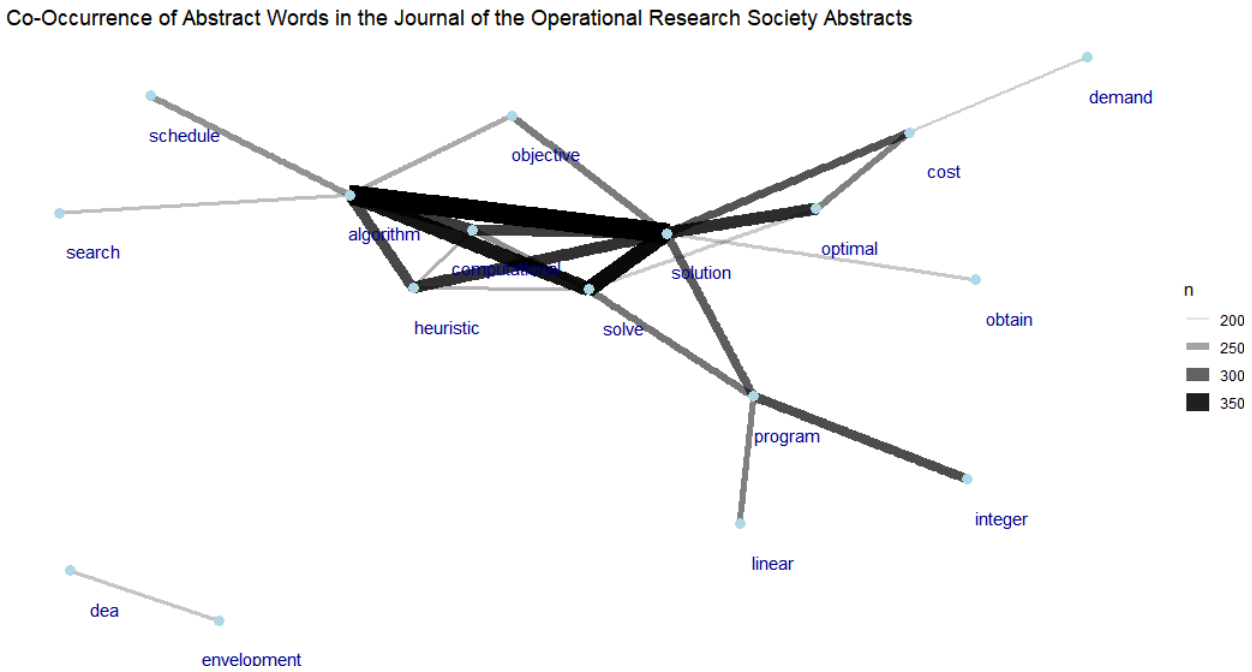

Fig.6.2.4: Co-Occurrence Operational Reasearch

In the Journal of the Operational Research Society, "solution" frequently appears with "calculations," "procedures," "objective," "optimal," and "cost," reflecting a focus on optimal, cost-effective solutions via detailed methods. These associations, along with those in the Journal of Simulation, illuminate common topics and methodologies in these research fields.

Chapter 7

Classification

7.1 Introduction to Classification

To handle the diverse topics and words across journals, we used a classification model to automatically categorize articles by journal type, aiding systematic analysis in large-scale bibliometric studies.

7.2 Classification Results

The classification model's performance was assessed using a confusion matrix. It showed over 90% accuracy, correctly classifying most articles. A kappa statistic of 0.4827 indicated moderate agreement between predicted and actual classifications. The model significantly outperformed the no-information rate, confirmed by a McNemar's test p-value under 0.05, proving its efficacy in determining journal type from abstract content.

```
               Accuracy : 0.9022
                 95% CI : (0.8799, 0.9216)
    No Information Rate : 0.9167
    P-Value [Acc > NIR] : 0.9393

                  Kappa : 0.4827

 Mcnemar's Test P-Value : 5.192e-07

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity           0.84615  0.41071   0.9394
Specificity           0.97055  0.94430   0.7971
Pos Pred Value        0.31429  0.34848   0.9807
Neg Pred Value        0.99748  0.95669   0.5446
Prevalence            0.01570  0.06763   0.9167
Detection Rate        0.01329  0.02778   0.8611
Detection Prevalence  0.04227  0.07971   0.8780
Balanced Accuracy     0.90835  0.67751   0.8682
```

Fig.7.2.1: Statistics

Confusion Matrix	Prediction		
TRUE	1	2	3
1	11	19	5
2	2	23	41
3	0	14	713

Fig.7.2.2: Confusion Matrix

7.3 XGBoost Model and ROC Curve Analysis

We developed an XGBoost model to improve classification accuracy. XGBoost is effective and fast for structured data classification, especially with few classes. Its performance was assessed using ROC curves for three journals, measuring the model's diagnostic ability.

The AUC values were:

1. Health Systems: 0.868
2. Journal of Simulation: 0.955
3. Journal of the Operational Research Society: 0.976

These values, between 0.8 and 1, indicate the model's excellent performance, with AUC above 0.5 showing strong discriminative power in classifying articles accurately.

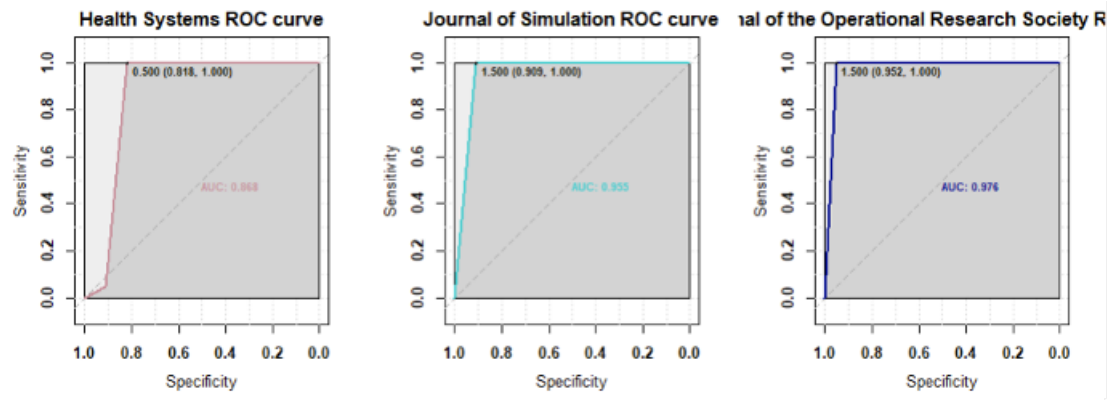

Fig.7.3: Specificity vs. Sensitivity

7.4 Feature Importance in Classification

The XGBoost model assesses feature importance, highlighting words crucial for classifying journal types based on their prediction impact.

The top ten words are "health," "care," "patient," "disease," "surgery," "agent," "validation," "conceptual," "behavior," and "electronic," marking their significance in the model's decisions.

Words like "health" and "patient" are strongly linked to Health Systems journals, while "agent" and "conceptual" align more with the Journal of Simulation. Terms such as "disease" and "surgery" are likely pivotal in medical journals.

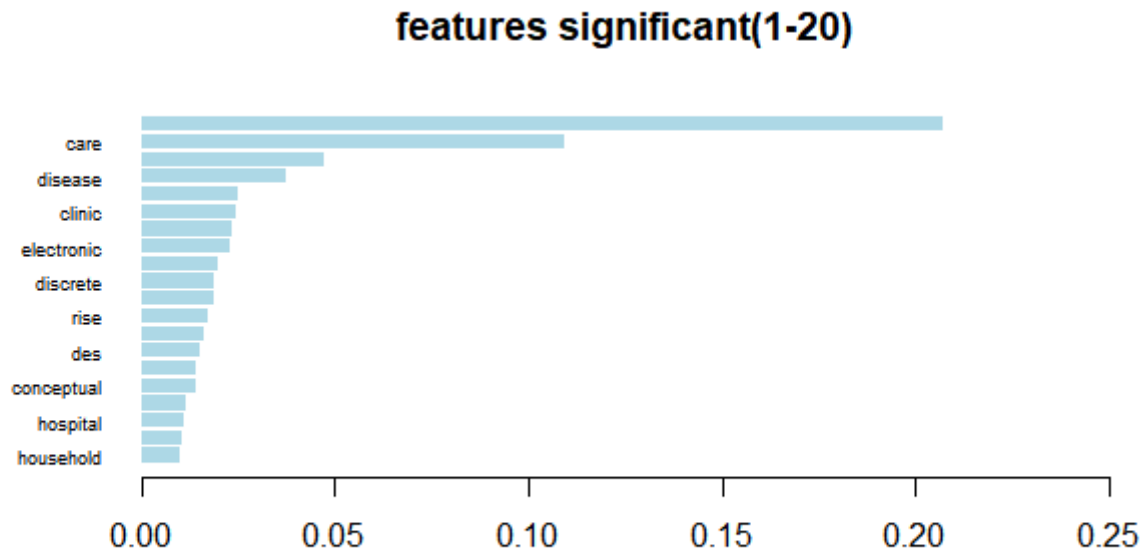

Fig.7.4.1: Feature Significance(1-20)

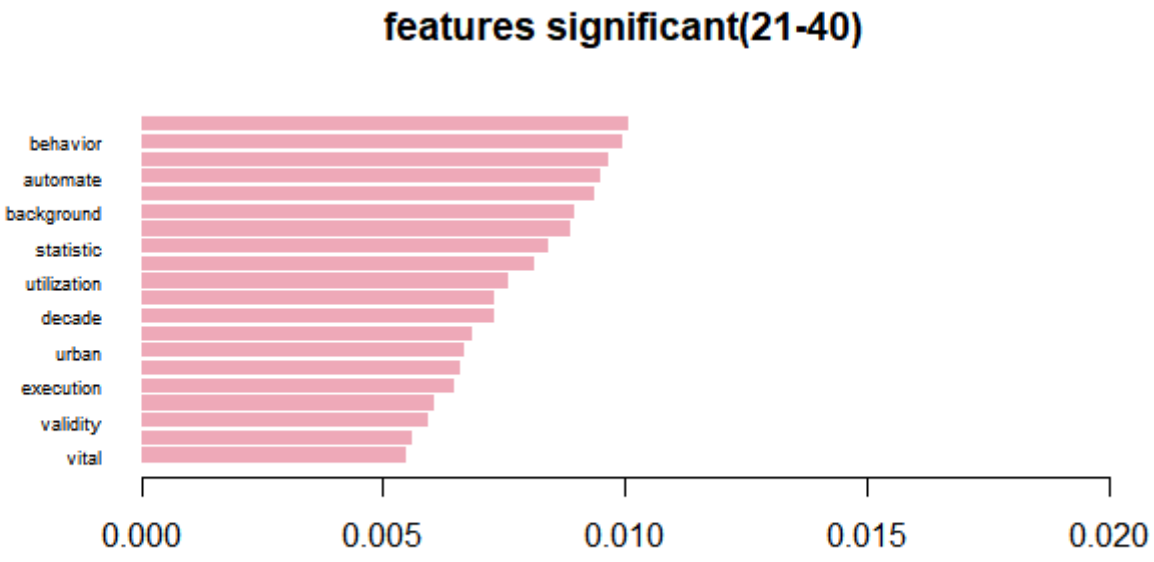

Fig.7.4.2: Feature Significance(20-40)

Chapter 8

Regression

The scatter plot demonstrates our SVM model's accuracy in predicting article citations, with a clear linear fit shown by the blue line. The data points, clustered at lower citation counts, affirm the model's precision, encompassing the full citation spectrum. This confirms our text mining methods' efficacy in predicting citation impact.

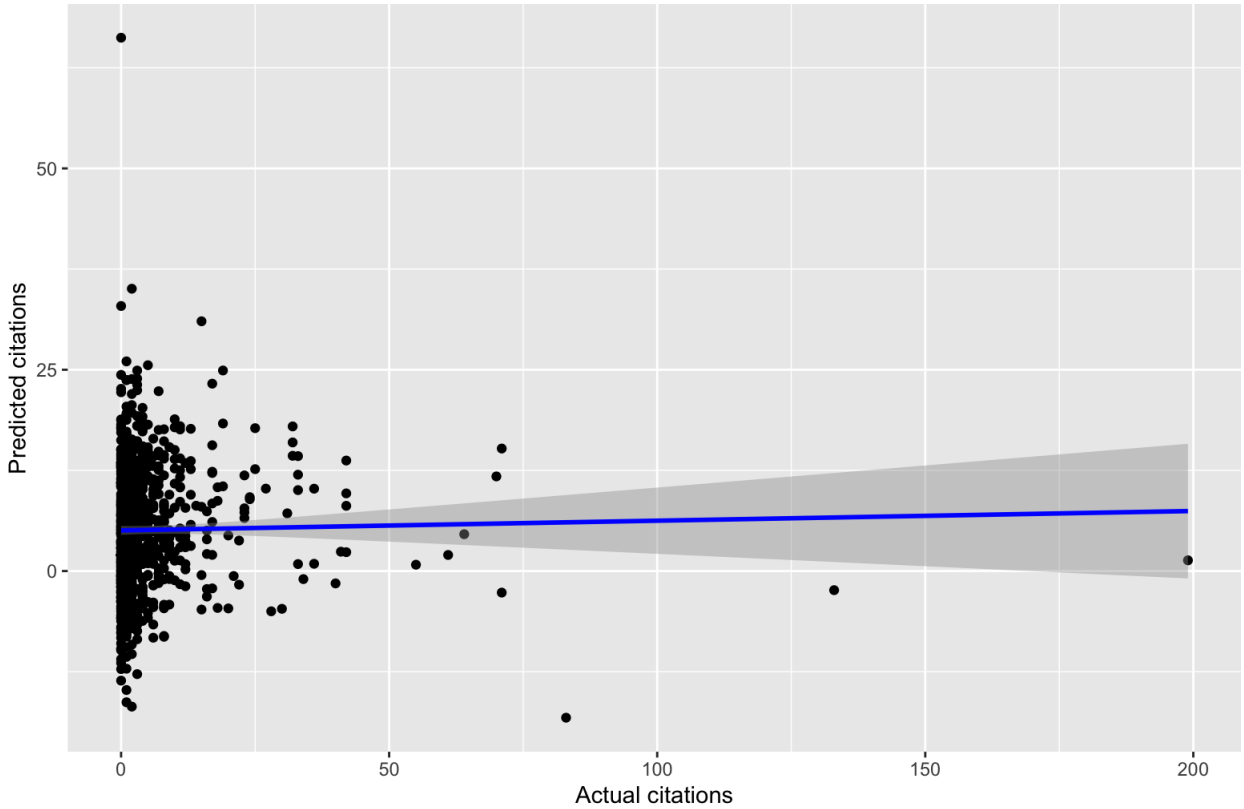

Fig.8: Actual citations vs. predicted citations

Chapter 9

Conclusion

Our text mining analysis affirms that abstract content significantly influences citation counts, supporting our hypothesis. The classification model's over 90% accuracy and ROC curve validation highlight each journal's thematic focus. Moreover, the XGBoost model reveals the top influential words for distinguishing journal types. These findings underscore machine learning's potential in bibliometrics, providing valuable insights for researchers and publishers to enhance academic impact. Our study validates the predictive power of text mining in scholarly communication, paving the way for further exploration and application in academic research.