EDUZONE: International Peer Reviewed/Refereed Multidisciplinary Journal (EIPRMJ), ISSN: 2319-5045
Volume 10, Issue 1, January-June, 2021, Impact Factor 7.687 Available online at: www.eduzonejournal.com

A Systematic Review of Research Trend Identification using Text Mining Techniques for Anomaly Detection

Pratik Badri

Independent Researcher

ABSTRACT

Text mining techniques have emerged as an important tool to detect anomalies and identify trends in research by analysing a large dataset of text. This includes academic papers, social media posts or other unstructured data. This research has conducted a systematic review of the application of text mining techniques in research based on both qualitative and quantitative data. The findings have highlighted the diverse application of text mining in different fields of research starting from science and healthcare to social computing and forest management.

Keywords: Text Mining, Trend Tracking, Social Computing, Anomaly, Publication Output, Text Semantics, NPL

INTRODUCTION

A. Background to the Study

Identifying and extracting aberrant features within information has been done for years using anomaly detection. A variety of methods have been employed to find anomalies. One of the primary methods in this field is text mining, that is becoming more and more relevant. Text mining frameworks which identify irregularities throughout their use are analysed in the current paper's Systematic Literature Review (SLR). Four approaches are used across the review to analyse the structures including anomaly detection uses, text mining strategies, text mining model effectiveness measures, as well as anomaly detection categorisation [1]. The aforementioned all have the ability to reveal research patterns in this age of swift advancement in technology.

B. Overview

The practice of gleaning significant knowledge and perspectives from semi-structured or unorganised textual information is known as text mining techniques. Methods for text mining have grown into an essential resource for academics and professionals across a variety of areas due to the rapid proliferation of digital information and the requirement to analyse enormous quantities of input [3]. Patterns, styles, and associations within textual information which would be challenging or unattainable to find via investigation by hand can be found with the aid of text mining.

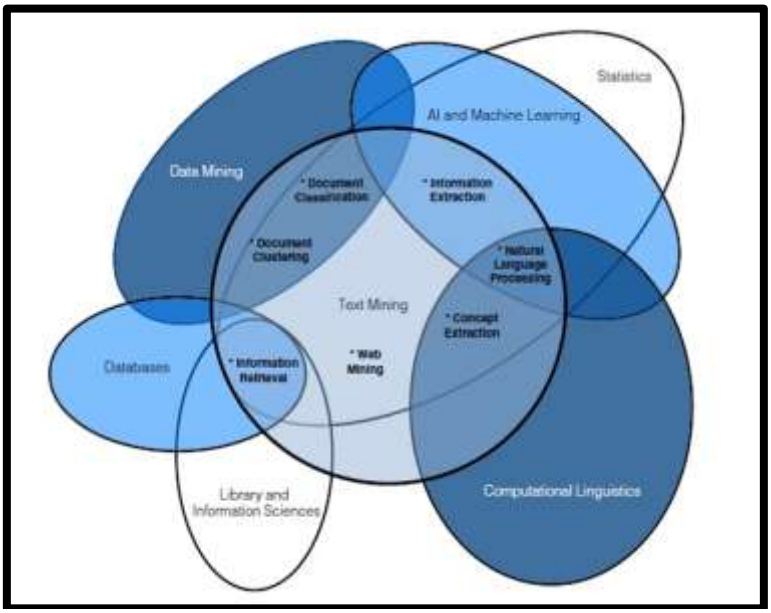

Figure 1: Venn Diagram of Text Mining

C. Problem Statement

Researchers now have the chance to learn more about forthcoming developments as well as subjects in their specific area of interest due to the growing accessibility of vast amounts of textual information through a variety of sources, including news stories, scientific papers, and social networking sites. However, subjective examination of this kind of information is frequently laborious and prone to individual limits, preconceptions, and mistakes [5]. By automatically collecting important details from text knowledge and offering an unbiased, data-driven strategy for evaluation, text mining algorithms can solve these difficulties.

D. Objectives

Finding and assessing text mining methods for identifying research patterns for anomaly detection is the primary goal of this paper.

- To categorise and distinguish amongst the many text mining methods for spotting research trends
- To evaluate various methods' advantages and disadvantages in terms of precision, effectiveness, flexibility, and generalisability
- To assess how well these methods detect patterns of research by contrasting the findings with the body of understanding and current literature
- To determine the obstacles and prospects for further study in this field

E. Scope and Significance

The application of text mining tools to find research patterns across a range of topics is examined in this systematic review. In order to detect developments in research, the investigation examines earlier studies regarding text mining techniques as well as their potential uses. Text mining instruments and techniques, sources of information and gathering approaches, and data analytics and visualisation strategies are all part of the research strategy and methodology. The research area found significant areas of study and trends in the academic panorama by analysing an array of publications from multiple disciplines. Additionally, the investigation contrasted its results with those of earlier research in the sector. The findings show how text mining methods might help guide funding choices and research goals.

LITERATURE REVIEW

A. A Brief Synopsis of Text Mining Methods along with Their Uses

Many fields, including medical care, business, social and behavioural sciences, and the study of nature, have made extensive use of text mining tools. These methods fall under three general areas, such as text evaluation, text visualisation, and text preparation. In order to increase the precision and effectiveness of evaluation, preparing text involves cleaning, normalising, and transforming original textual information [7]. Approaches for analysing texts may also be further divided into two categories such as trained and untrained approaches. Text categorisation and evaluation of sentiment are two examples of supervised learning approaches that need labelled training information. Unsupervised instruction comprises methods that lack labelled data, like topic modelling and text grouping. Networking graphs, word cloud visualisations, and topic mappings are examples of visual representations of text data created using text visualisation algorithms.

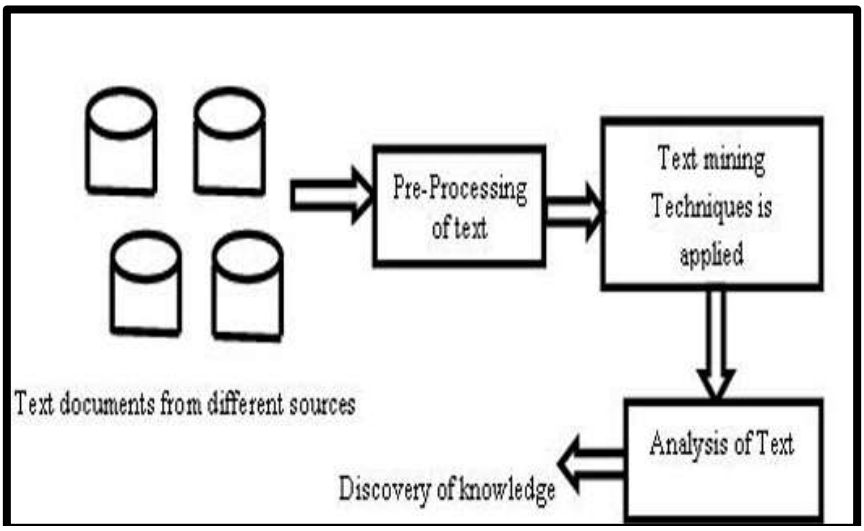

Figure 2: Process of Text mining

B. Evaluation of Present Studies and Information Gaps Critically

Notwithstanding the prospective advantages of text mining methods for spotting study patterns, there are a number of restrictions and difficulties that must be resolved. The accessibility and standard of textual information are two major drawbacks. The reliability and precision of outcomes may be impacted by the arbitrary, unclear, and chaotic nature of text-based information [2]. Output may also be impacted by the accessible nature of written material from other sources. Text information gathered from scientific papers could differ through written content from online forums, for instance. The choice of suitable text mining methods for spotting research patterns is another difficulty. The success of various approaches may vary depending on the goals and circumstances surrounding the study, and each may have unique advantages and disadvantages [8]. As a result, scholars have to carefully assess whether methods involving text mining are appropriate for a certain study subject.

C. Text Anomaly Identification: Principles and Uses

Text anomaly detection, sometimes referred to as uniqueness or aberration exploration, finds information in text which dramatically departs beyond the norm. Efficient fraud detection, information security, online communication analysis, including healthcare data management all depend on this identifying procedure. Obtaining textual elements that are essential for identifying anomalies, including word frequency ranges, syntactic trends, along with contextual links, is a fundamental part of TAD (Text anomaly detection) [4]. The capacity of systems for identifying anomalies has significantly increased with the development of text mining techniques, especially through machine learning and deep learning.

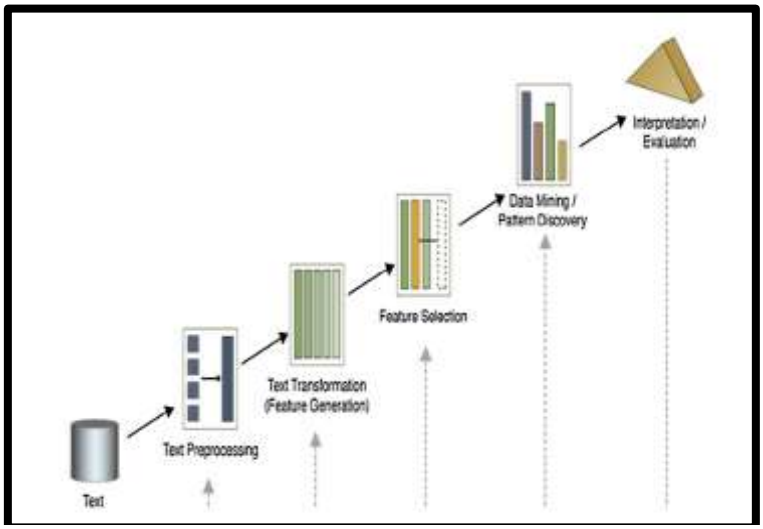

Figure 3: Scope of text mining in anomaly identification

D. Future Direction of Text Anomaly Detection Via Technological-Aid

Recently, there has been a lot of interest in using text mining to find anomalies in literature. at contrast to more conventional deep learning techniques, it is successful at identifying connections between contexts as well as long-range dependencies throughout textual information, enhancing efficiency [6]. Text mining outperforms other methods, such as embedded words, in terms of comprehending context and the significance of words, according to the prior review. This effectively detects semantic domain redundancy, improving anomaly detection across the range of real-world technology implementation.

METHODOLOGY

A. Research Design

The research design that was chosen to conduct this systematic review on research trend identification through text mining techniques is the exploratory design. The exploratory design is mostly used for research on concepts that have little prior knowledge [9]. Hence, this design has allowed gaining deep insights into anomaly detection through text mining from previous studies. The adaptability and flexibility of the exploratory design have allowed refining the research questions to delve deeper into the research topic. The cause and effect understanding that explanatory design offers were suitable to determine the reasons behind anomalies occurring in research.

B. Data Collection

Following a mixed-method approach, this research uses both secondary quantitative and qualitative data to determine the text mining techniques in trend identification in research. Qualitative data was obtained from academic journals,

case studies and other relevant literary materials to obtain rich insights on anomaly detection in research through text-mining techniques. On the other hand, quantitative data was collected from existing research studies, charts or statistical analysis highlighting research trend identification through text mining. The advantage of secondary data collection over primary is the time-saving aspect [10]. As data was obtained from already existing literature research on trend identification, resources and time were saved, which would otherwise the wasted in primary data collection. Moreover, secondary data enhances the integrity and credibility of findings as the chances of manipulated information are limited.

C. Case Studies/Examples

Case Study 1: Trend tracking through text-mining approaches in research on forest management

The case study mentions that to date, a bulk number of studies were located on searching for research on forest management in databases. Considering how unfeasible it is to examine this bulk number of publications and verify which topics are important or trending, the most scientific and effective method to derive information would be text mining. Text mining was used in this case study to reveal in which aspect the research on forest fire concentrates, which study topics have emerged or how the level of importance of chosen issues has changed over the years [11]. The important concepts that have emerged through this analysis include damage to vegetation, affected species, fire management, post-fire actions and structural changes among others.

Case Study 2: Identifying trends in social computing through text-mining

By applying the "rule-based anomaly detector" WSARE, research trend analysis was conducted in this case study for social computing research. From several articles collected from two databases, a wide range of social computing studies were identified, among which the majority were based on engineering and computer science topics. The countries which were the largest contributors to the studies were also identified. On an application of the "interdisciplinary network evaluation analysis", the close collaboration between subject categories was explored as well. The case study concluded that anomaly detection models had a high potential to identify hidden research trends, which makes them a useful tool in forecasting [12].

D. Evaluation Metrics

Table 1: Evaluation Metrics

Metric	Description	Purpose
Precision	Measures the fraction of anomalies detected from the actual number of anomalies existing.	Assesses the accuracy of anomaly identification [13].
Recall	To detect the correctly identified anomalies from the actual number of anomalies.	This metric is used to measure the ability of a model to identify all the relevant anomalies without missing any true anomalies [14].
ROC-AUC	To evaluate the performance of classifiers across different thresholds against the overall performance.	The ability of a model to distinguish between anomalies and normal instances is evaluated through this metric [15].
F1-score	The harmonic mean between precision and recall generates a balance between the two metrics.	Providing a balanced measure of the performance of a model, especially when imbalanced datasets have to be dealt with [16].

(Source: Self-Developed)

RESULTS

A. Data Presentation

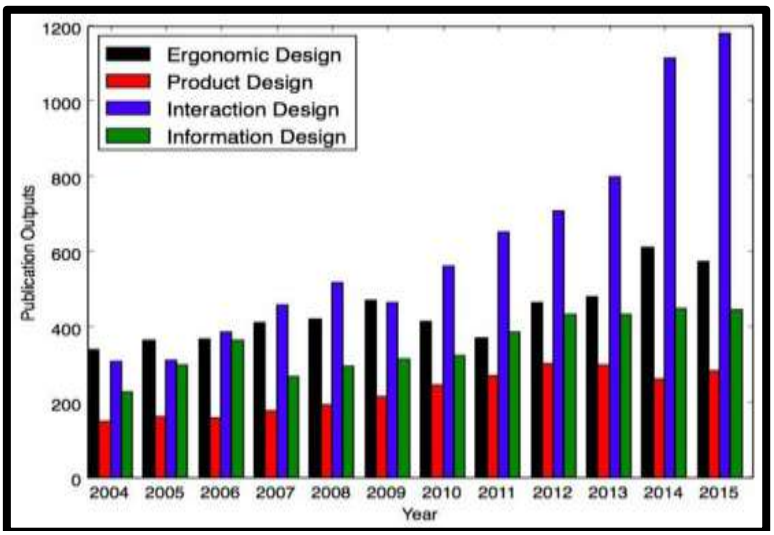

Figure 4: Publication Outputs of Academic Branches Identified Through Text Mining In Design Research

The graph in Figure 4 above demonstrates the time trend analysis of research through text-mining techniques. The figure displays the trends about citation and publication outputs that were obtained in design research through text-mining. Starting from a mere number of around 300 in 2004, the number of studies on interaction design has increased to nearly 1200 by 2015. The product design also started with less than 200 publication outputs during 2004 but by 2015, the number crossed the 200 mark. The number of studies on ergonomic design has fluctuated but has rapidly surged and exceeded 600 in 2014.

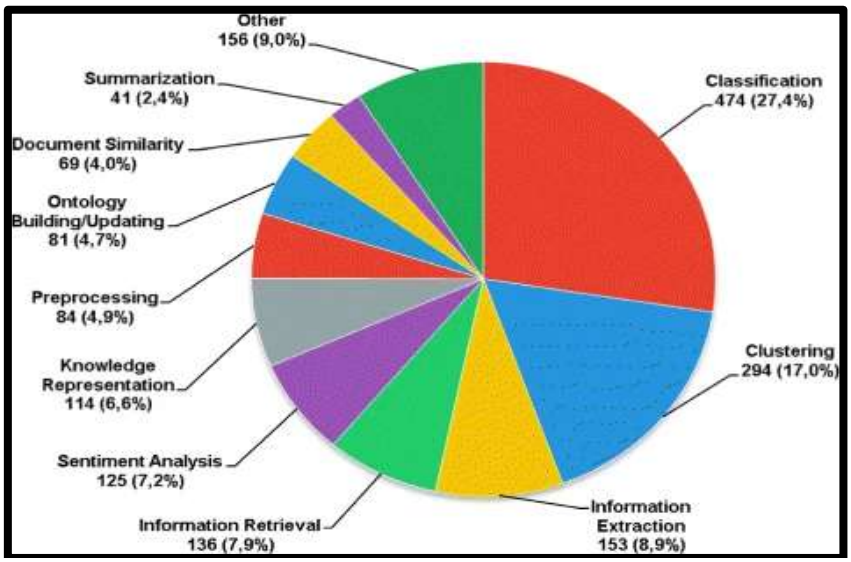

Figure 5: Text mining tasks identified from the literature through text mining

Figure 5 above has answered the question of which text mining tasks are most considered in text semantics. The distribution of different text-mining tasks presented in the literature mapping in the figure above reveals clustering and classification as the most frequent tasks. The classification task in text mining was used 474 times by researchers, followed by clustering, which was used 294 times. Tasks like summarisation and document similarity constitute only a minimal section of the pie chart.

B. Findings

A significant surge was observed in publication output associated with interaction design, indicating the level to which the popularity of the internet has influenced the rapid development of this design [17]. The graph for product design has

also shown rapid growth, except for some temporary interruptions in its development. However, compared to publication outputs about interaction design, other design research has not experienced much progress over the years. The findings also highlight the importance of text-mining techniques in identifying major academic trends and branches in design research. As these are basic tasks in text-mining, they form the groundwork for other text-mining tasks like ontology building or sentiment analysis. Hence, this data points out the important role of text semantics in text mining. Addressing the lack of adequate studies integrating different branches of research, these findings give a general summary of the areas lacking development in research [18]. This can serve as a guiding framework for researchers willing to work with semantics-related text mining.

C. Case Study Outcomes

Table 2: Case Studies Key Outcomes

Case Study	Key outcomes
Case Study 1: Trend tracking through text-mining approaches in research on forest management	Research regarding "sustainable forest management (SFM)" is specific to contexts and developing over time but the changes are rarely investigated. The most convergent narrative within the research landscape of SFM indicates that SFM research has evolved through three phases [11]. The early phases cover issues regarding land usage while the second phase focuses on forestry and the third phase experiences a publication downturn and the focus shifts towards climate change.
Case Study 2: Identifying trends in social computing through text-mining	According to the anomalies detected by the text-mining technique WSARE, the research on social computing has gradually shifted from the traditional fields such as engineering and computer science towards the fields of health, medicine and communication. A wide range of new subjects have emerged in recent years including crowd-sourcing, sentimental analysis and e-health [12]. These anomaly detection models can be highly used in forecasting purposes in other fields of research.

(**Source: Self-Developed)**

D. Comparative Analysis of Literature Review

Table 3: Comparative Analysis of Literature

Author	Focus	Key Findings	Literature Gap
[2]	Identifying the most fundamental techniques and tasks in text mining.	A large number of growing online scientific articles has enabled researchers to access scientific information easily but identifying pertinent articles can be difficult without the application of text-mining techniques.	The limited transparency on the research methodology is a limitation of this study.
[3]	Overviewing the pre-processing techniques involved in text mining.	Text mining is among the research-related areas of data mining, which is used in different research domains like NLP, text clustering, classification and information retrieval.	Limited discussion on the challenges of different pre-processing text mining techniques.
[4]	Overview of the concept of qualitative analysis.	Thematic analysis is a common method in "qualitative content analysis "(QCA)" involving the formation of categories from text obtained from the work of other scholars.	The concept of text mining is not emphasised enough.

[6]	Identifying the selection criteria involved in text mining.	General criteria like comprehensiveness, flexibility and usability, achievement criteria, data analysis criteria and other miscellaneous criteria can be involved in measuring the effectiveness of a text-mining technique.	The basis for proposing the evaluation criteria was not mentioned.
[7]	Evaluation of the text-mining techniques, issues and applications.	Information extraction, information retrieval and NLP are the types of text-mining techniques and applications found in the field of digital library, research, life science and social media among others.	Poor transparency about research methodology.
[8]	Evaluating the application of text-mining in information systems research.	Probabilistic topic modelling is an unsupervised technique in text mining that can be used to explain user satisfaction with IT artefacts.	The limited effort to validate the reliability of the information provided.

(Source: Self-Developed)

A small summary of previous literature was provided in Table 3 above, highlighting the types of text mining and their applications. The gaps in previous literature were also outlined.

DISCUSSION

A. Interpretation of Results

The findings point out the huge application of text mining techniques in trend identification of research, starting from social computing to forest management studies. An exploration of the publication outputs regarding different academic branches identified with text mining, interaction design was found to be the most identified branch. In terms of time trend analysis, classification and clustering were reported as the most frequently used text mining tasks among researchers [18]. The qualitative findings reveal that the growing number of online scientific articles justify the need for text-mining techniques to identify relevant information. Text clustering, classification and NLP are different techniques in text mining and their applicability is diverse.

B. Practical Implications

The primary practical implication of text mining techniques in the identification of research trends is in the discovery of efficient knowledge. The retrieval of relevant information from a bulk number of sources would not just be time saving for researchers in different fields but also help in making informed decisions and allocating resources efficiently both researchers and business organisations can make strategic choices through text mining [19]. Efficiently identifying the anomalies in research through text mining techniques will allow researchers to structure the research paper and organise it properly, making it easier for others to access information. Both time and resources would be saved which can be utilised for in-depth analysis.

C. Challenges and Limitations

The primary challenge in this research was regarding the mixed-method approach which was both time-consuming and complex. Moreover, since secondary data was relied upon, finding sufficient and relevant information on text-mining techniques was a challenge. One limitation of the study is that the findings were not specific to any field of research and hence, are too generic.

D. Recommendations

To ensure the anomaly detection through text mining techniques is effective, the focus needs to be on a robust text pre-processing [20]. This would involve cleaning the text for consistent analysis, tokenisation or splitting into individual words and extraction of key features.

CONCLUSION AND FUTURE WORK

The research has evaluated that due to the increasing number of scientific articles in different fields, the role of text-mining techniques has become more important in diverse fields. The model selection and data preparation aspects of text-mining can be explored further in future research. Moreover, the criteria involved in ensuring that the chosen text-mining technique is appropriate for a specific research area and is suitable to the characteristics of collected data can

also be identified. Additionally, unsupervised techniques like "Latent Dirichlet Allocation (LDA)" and algorithms like hierarchical clustering and k-means are also some underexplored areas that could be focused in future research.

REFERENCES

[1]. A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou, "Using text mining for study identification in systematic reviews: a systematic review of current approaches," Systematic Reviews, vol. 4, no. 1, Jan. 2015, doi: https://doi.org/10.1186/2046-4053-4-5.

[2]. M. Allahyari et al., "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," arXiv:1707.02919 [cs], Jul. 2017, Available: https://arxiv.org/abs/1707.02919

[3]. S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining-an overview," nternational Journal of Computer Science & Communication Networks , vol. 5, no. 1, pp. 7–16, 2015.

[4]. U. Kuckartz, "Qualitative text analysis: A systematic approach," pp. 181–197, 2019.

[5]. R. Alghamdi and K. Alfalqi, "A Survey of Topic Modeling in Text Mining," International Journal of Advanced Computer Science and Applications, vol. 6, no. 1, 2015, doi: https://doi.org/10.14569/ijacsa.2015.060121.

[6]. H. Hashimi, A. Hafez, and H. Mathkour, "Selection criteria for text mining approaches," Computers in Human Behavior, vol. 51, pp. 729–733, Oct. 2015, doi: https://doi.org/10.1016/j.chb.2014.10.062.

[7]. R. Talib, M. Kashif, S. Ayesha, and F. Fatima, "Text Mining: Techniques, Applications and Issues," International Journal of Advanced Computer Science and Applications, vol. 7, no. 11, 2016, doi: https://doi.org/10.14569/ijacsa.2016.071153.

[8]. S. Debortoli, O. Müller, I. Junglas, and J. Vom Brocke, "Text mining for information systems researchers: An annotated topic modeling tutorial," Communications of the Association for Information Systems (CAIS) , vol. 39, no. 1, p. 7, 2016.

[9]. B. Rittle-Johnson, J. R. Star, and K. Durkin, "The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving.," Journal of Educational Psychology, vol. 101, no. 4, pp. 836–852, Nov. 2009, doi: https://doi.org/10.1037/a0016026.

[10]. A. Schober, N. Šimunović, A. Darabant, and T. Stern, "Identifying sustainable forest management research narratives: a text mining approach," Journal of Sustainable Forestry, vol. 37, no. 6, pp. 537–554, Feb. 2018, doi: https://doi.org/10.1080/10549811.2018.1437451.

[11]. Q. Cheng, X. Lu, Z. Liu, and J. Huang, "Mining research trends with anomaly detection models: the case of social computing research," Scientometrics, vol. 103, no. 2, pp. 453–469, Mar. 2015, doi: https://doi.org/10.1007/s11192-015-1559-9.

[12]. A. Guille and C. Favre, "Event detection, tracking, and visualization in Twitter: a mention-anomaly-based approach," Social Network Analysis and Mining, vol. 5, no. 1, May 2015, doi: https://doi.org/10.1007/s13278-015-0258-0.

[13]. C. C. Aggarwal, "An Introduction to Outlier Analysis," Outlier Analysis, vol. 1, no. 1, pp. 1–34, Dec. 2016, doi: https://doi.org/10.1007/978-3-319-47578-3_1.

[14]. B. Nie and S. Sun, "Using Text Mining Techniques to Identify Research Trends: A Case Study of Design Research," Applied Sciences, vol. 7, no. 4, p. 401, Apr. 2017, doi: https://doi.org/10.3390/app7040401.

[15]. R. A. Sinoara, J. Antunes, and S. O. Rezende, "Text mining and semantics: a systematic mapping study," Journal of the Brazilian Computer Society, vol. 23, no. 1, Jun. 2017, doi: https://doi.org/10.1186/s13173-017-0058-7.

[16]. A. Mahapatra, N. Srivastava, and J. Srivastava, "Contextual Anomaly Detection in Text Data," Algorithms, vol. 5, no. 4, pp. 469–489, Oct. 2012, doi: https://doi.org/10.3390/a5040469.

[17]. Chintale, P. (2020). Designing a secure self-onboarding system for internet customers using Google cloud SaaS framework. IJAR, 6(5), 482-487.

[18]. INNOVATIONS IN AZURE MICROSERVICES FOR DEVELOPING SCALABLE", int. J. Eng. Res. Sci. Tech., vol. 17, no. 2, pp. 76–85, May 2021, doi: 10.62643/

[19]. "Balancing Privacy and Utility: Anonymisation Techniques for E-commerce Logistics Data", int. J. Eng. Res. Sci. Tech., vol. 17, no. 2, pp. 65–75, Apr. 2021, doi: 10.62643/.

[20]. "Intelligent Process Automation in S/4 HANA FICO: A Machine Learning Approach ", IJIEE, vol. 10, no. 2, pp. 57–70, Feb. 2020, doi: 10.48047/aqtbk646.