





A Review Of Text Mining Techniques: Trends, and Applications In Various Domains

Hiba J. Aleqabie^{1*}, Mais Saad Sfoq², Rand Abdulwahid Albeer¹, Enaam Hadi Abd¹

¹Computer Science Department, Collage Of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq

²Information Technology Department, Collage Of Computer Science and Information Technology University of Kerbala., Karbala, Iraq

*Corresponded Author: * Hiba J. Aleqabie

DOI: <https://doi.org/10.52866/ijcsm.2024.05.01.009>

Received June 2023 ; Accepted September 2023 ; Available online January 2024

ABSTRACT

Text mining, a subfield of natural language processing (NLP), has received considerable attention in recent years due to its ability to extract valuable insights from large volumes of unstructured textual data. This review aims to provide a comprehensive evaluation of the applicability of text mining techniques across various domains and industries. The review starts off with a dialogue of the basic ideas and methodologies that are concerned with textual content mining together with preprocessing, feature extraction, and machine learning algorithms. Furthermore, this survey highlights the challenges faced at some stage in implementing textual content mining strategies. Additionally, the review explores emerging tendencies and possibilities in text-mining research. It discusses advancements in deep learning models for text evaluation, integration with different AI technologies like image or speech recognition for multimodal analysis, utilization of domain-unique ontologies or information graphs for more desirable information of textual facts, and incorporation of explainable AI strategies to improve interpretability. The findings from this overview are analyzed to identify common developments and patterns in text mining packages across extraordinary domain names. The consequences of this paper will advantage researchers by means of imparting updated expertise of modern practices in textual content mining. Additionally, it will manual practitioners in selecting suitable strategies for their unique application domain names while addressing capacity-demanding situations.

Keywords: Text Mining; Natural Language Processing; Challenges; Opportunities.

1. INTRODUCTION

The number of digital written information continues to grow, with books being digitized and archived for future generations. Text mining techniques are increasingly being utilized to extract information automatically from this enormous body of text. These methods are used to filter through papers, assess them, and find insights[1].

The exponential growth of digital data in recent years has produced an urgent need for effective techniques to extract useful insights from unstructured text. Text mining, also known as text analytics or natural language processing (NLP), has emerged as a powerful method for revealing the enormous value hidden within textual data. This emerging subject encompasses a diverse set of techniques and procedures for gathering, analyzing, and analyzing data from a multitude of sources. Social media, biological literature, customer input, and other sources are examples. Text mining is used in a variety of fields, including business, healthcare, social media analysis, and scientific study. As businesses attempt to capitalize on the latent information hidden in unstructured text, text mining has gained traction for its capacity to unearth actionable insights and enhance data-driven decision-making. Recent research has shown that text mining may be used to solve complicated problems in a variety of disciplines, demonstrating its potential to revolutionize information extraction and knowledge discovery. For instance, research [2] showcased the utility of advanced text mining techniques in analyzing social media data to infer public sentiment and forecast market trends, while the work of [3] illustrated the application of text mining in healthcare for enhancing clinical decision support and biomedical research. These recent references underscore the increasing relevance and impact of text mining, setting the stage for further exploration and innovation in this dynamic and rapidly evolving field.

Furthermore, recent studies have demonstrated the effectiveness of text mining techniques in various applications. For instance, a study by [4] explored the use of text mining algorithms in analyzing social media data to predict stock market trends. The results showed that sentiment analysis combined with other textual features can provide valuable insights into stock market movements. Another study by [5], focused on applying text mining techniques to analyze online reviews for product recommendation systems. The research highlighted the importance of sentiment analysis and topic modeling in improving the accuracy of product recommendations.

Text mining is garnering prominence among academics, and its application is expanding exponentially in many fields of study. Standard applications for text mining may include text summarization [6], information retrieval [7], information clustering [8], text categorization [9], information extraction [10], language identification [11], and phrase structure identification [12].

In text summarization different researches were published for instance, The study [13] covers abstractive text summarization, its drawbacks, and the use of transfer learning and deep reinforcement learning. It examines the benefits and drawbacks of reinforcement learning techniques such as sequence-level training and policy gradient methods. The significance of transfer learning in enhancing performance across many disciplines and languages is also examined in this research. In addition, it examines current studies and talks about difficulties and unanswered problems in this quickly developing subject. In [14] reviews recent approaches for abstractive text summarisation using deep learning models, focusing on Gigaword and CNN/Daily Mail datasets. It discusses the quality of summarisation using Recall-Oriented Understudy for Gisting Evaluation 1 (ROUGE1), ROUGE2, and ROUGE-L metrics. Recurrent neural networks with attention mechanisms and long short-term memory (LSTM) are the most prevalent techniques. Pretrained encoder models achieve the highest ROUGE1 and ROUGE-L values, but face challenges like golden token unavailability, out-of-vocabulary words, and fake facts. In [6] Automatic summarization is a method that condenses lengthy passages of content to swiftly provide important information. Summarization may be used in both text and video, but it uses a distinct approach to convey the topic's essence. This research applies automated text summation using natural language processing to YouTube videos by transcription and application of the study's summary phases. Using the term frequency-inverse document frequency (TF-IDF) approach, the text's words and phrases were counted to determine which significant. In the information retrieval field, several instances can be illustrated. In paper [15] suggests using a question-answering approach to automatically extract textual data on infrastructure damage. The technique was trained using 143 reports from the National Hurricane Center and makes use of bidirectional encoder representations from transformers. The hurricane and earthquake datasets yielded F1-scores of 90.5% and 83.6%, respectively, for the model. This research [16] integrates sparse and dense approaches to give a conceptual framework for natural language processing and information retrieval. By dividing the fundamental text retrieval problem into a logical scoring model and a physical retrieval model, it suggests a representational strategy. The framework proposes open research questions and lists several retrieval techniques. Additionally, it links the framework to information access and natural language processing tasks including sentence similarity. In This study [16] explores how using Wikipedia concepts in query context can improve proactive information retrieval on noisy text. Two models use entity linking to associate topics with relevance, and experiments show clear relevance signals in Wikipedia concepts. Wikifying the query context can disambiguate meaning, further aiding proactive retrieval. In the field of text categorization an instance may be illustrated. The study [17] investigates text classification in relation to document and social media consumption. It concentrates on topic-based feature extraction and selection, employing the Uncapacitated P-Median Problem (UPMP) for Twitter data clustering. To address the UPMP issue, a brand-new hybrid genetic bat algorithm (HGBA) is put forth. The study evaluates victims' needs both during and after a tragedy using Twitter. Tests conducted on the OR-Library dataset demonstrate that the suggested method effectively identifies themes and classifies text. Predicate-argument information is incorporated into the bi-encoder technique for paraphrase detection used in this work [18]. Experiments demonstrate that the suggested model performs substantially better than SBERT/SRoBERTa with very minor parameter adjustments. Performance is greatly improved by the predicate-argument-based component, which outperforms cross-encoders and cross-encoders. The study [19] emphasizes developments in reading science, such as word recognition, comprehension, and universal viewpoints. It also emphasizes how different models, theories, and evidence-based approaches may improve teaching tactics.

Also, it involves different domains applications; these domains are explored below.

1.1 Healthcare

Text mining is extensively used in healthcare for clinical decision support, disease surveillance, and pharmacovigilance tasks. For instance, [20] has used text mining methods to extract adverse drug occurrences from electronic health records (EHRs), enabling early detection of potential drug-related risks. Another study by [21][22] utilized text mining to identify patterns of suicidal behavior in clinical notes of psychiatric patients. Among the other things is the utilization of mobile healthcare (mHealth) apps, which are mobile device-based platforms that enable healthcare providers—such as physicians, pharmacists, hospitals, and clinical laboratories—to communicate with and

gather a large number of reviews and responses from consumers—that is, patients—. Subsequently, data mining and natural text processing techniques are used to evaluate the reviews and responses to determine data polarity and consumer satisfaction [23]. Article [24] shows how text mining and natural language processing may be used to retrieve significant information from healthcare procurement data. It emphasizes the significance of employing sophisticated analytics approaches to extract insights from unstructured textual data for better healthcare decision-making. The importance of electronic medical records (EMRs) in the healthcare industry is first discussed in the article[25]. EMRs have a lot of important data that can be used for research, decision-making, and healthcare enhancement. However, manual analysis becomes problematic due to the massive volume of data contained in EMRs; therefore, automated data processing and text-mining techniques must be used in investigating several text mining techniques such as sentiment analysis, information extraction, and natural language processing (NLP). So the study [26] highlights how text mining can effectively analyze and understand clinical medical information, leading to better healthcare results. The article[27] introduces HIMERA, a semantically annotated corpus, and a time-sensitive terminological inventory for text mining in 19th- and 20th-century medical texts. It shows their effectiveness in detecting historical term relationships and introduces a TM pipeline for efficient exploration and search.

This thesis investigates in[28] the degree of accuracy necessary for text mining methods to analyze clinical outcomes for medical research. It emphasizes the need to maintain high accuracy in text mining while working with sensitive medical data. In order to reduce mistakes and inconsistencies in the extracted data, it highlights the necessity of using strong methodology and validation procedures, indicating that to increase the accuracy of text mining tools, a cooperative strategy including domain experts, data scientists, and physicians is necessary. In order to better understand text mining for radiology reports, the study[29] first examines a variety of approaches and algorithms, such as information retrieval, machine learning, and natural language processing (NLP). It explores how well they extract pertinent data from radiological reports and how they can do it better, it highlights how text-mining algorithms may help radiologists to make evidence-based decisions, discover and prioritize significant results, and improve patient care. The automated classification of positive and negative healthcare phrases in sailors' textual healthcare documents is investigated in this study [30] using lexicon sentimental analysis. This resulted from an absence of computationally assisted experimental assessments. The LASSO regression technique is used to examine these text documents and categorize illnesses and their accompanying symptoms. Analyzing TF-IDF measurements might result in a display of the frequency of symptomatic data for each condition. Cardiologists employ digital technologies to reach patients, doctors, and the general public. They perform a variety of tasks, such as documenting cardiovascular parameters, helping with diagnosis, educating patients, and instructing laypeople in cardiopulmonary resuscitation. Health care has already benefited much from this profession, and we expect it to continue growing. In [31] , the researcher identified and analyzed the academic literature on the use of digital technology in cardiology using a bibliometric technique, revealing popular research subjects, important authors, institutions, nations, and journals. We have included the cardiovascular diseases and diagnostic instruments that are most frequently looked at in this discipline.

Using text mining tools in [32]to investigate the developments and patterns in the field of Alzheimer's disease research, the study's main objective is to examine scientific literature in order to learn more about how this field of study has developed. The researchers were able to pinpoint important subjects, well-known writers, and new directions in the field of Alzheimer's disease research through text mining techniques. From a health technology standpoint, this method offers useful information on the state and trajectory of Alzheimer's disease research.

1.2 Social Media Analysis

With the proliferation of social media platforms like X (Twitter)and Facebook, text-mining techniques have been employed to analyze user-generated content for sentiment analysis, opinion mining, and trend detection. For instance, [33] conducted a study using Twitter data to predict stock market activities based on sentiment analysis of tweets related to specific companies or financial terms. The study [34] aims to develop a tool that can automatically analyze the sentiment expressed by users towards fashion images on Instagram. Sentiment analysis involves identifying and classifying emotions and opinions expressed in textual data.

Due to the rise of fake news—which poses a threat to social cohesion and trust, fostering political polarization and distrust— and the vast amount of news disseminated through social media, automatic systems for fake news detection have been developed[35]. Sentiment analysis, a part of text analytics, is used to determine the polarity and strength of sentiment in fake news detection approaches. Future requirements include multilingualism, explain ability, bias mitigation, and multimedia treatment.

This study[36] proposes a text mining technique for online travel evaluations that makes use of text categorization and natural language processing technologies. In order to increase efficiency, it examines the reliability of internet review content. Text classification technology and sentiment analysis are used in this method's evaluation of hotels and picturesque sites. The algorithm mines service features and finds new terms based on left and right entropy, as well as mutual information. Examining the user reviews and applying Latent Dirichlet Allocation (LDA) to comprehend movie subjects, the system [36] in makes movie recommendations to users. The general sentiment linked with each movie is ascertained by the algorithm through the collection of user comments and the analysis of sentiment. Based on user choices, LDA creates individualized recommendations by classifying movies into distinct subjects.

The study [37] analyzes Twitter tweets about vegan food, revealing that ethical, personal health, and environmental factors are the main drivers for vegan food choice. However, there is a limited number of sustainability-motivated tweets. Instead, value propositions about personal health attributes and consumption benefits are more appealing. The polarity of attitudes between vegans and non-vegans suggests that a single value proposition may not reach both groups simultaneously.

The study[38] analyzed 8229 hotel reviews from December 2019 to July 2021, focusing on fundamental selection attributes and their association with customer satisfaction. Results showed that Service and Dining factors significantly affected customer satisfaction, emphasizing the need for specific services, especially after COVID-19. Understanding online reviews can help develop sustainable strategies for the hotel industry, ensuring customer satisfaction and repurchase intention.

Nonpharmaceutical treatments are required because the COVID-19 pandemic has exposed transportation hubs to threats to public health. But not all of us are aware of how building design affects the effectiveness of policies. The study [39], which examined 103,428 Google Maps evaluations of 64 US hub airports, discovered that although staff and shops were well rated, service and space received indifferent or negative ratings. The project seeks to increase transportation hubs' capacity to withstand health emergencies in the future.

1.3 Customer Relationship Management

Text mining has been applied in customer relationship management (CRM) to extract insights from customer feedback, reviews, and surveys.[40] developed a text mining framework to analyze online customer reviews and identify critical factors influencing customers' purchase decisions.[41] seeks to create a sentiment analysis model that successfully categorizes movie reviews as positive or negative, depending on the sentiment conveyed. The authors propose a novel approach that combines word embedding techniques with semantic orientation to enhance sentiment classification accuracy. [42] This research examines the impact of customer relationship management (CRM) on customer satisfaction in private sector organizations. It focuses on the relationship between customer satisfaction and CRM, along with the variables influencing it. The study uses a quantitative approach to analyze responses from participants. The results show that CRM significantly impacts customer satisfaction, with the research question answering participants' responses. In this chapter[43] it discusses dialogue management using discourse, introducing an imaginary discourse tree for on-demand background knowledge and a lattice walk approach. It also introduces the Doc2Dialogue algorithm, which converts text into hypothetical dialogues based on discourse tree analysis, extending chatbot training datasets. The algorithm's deployment is crucial for successful chatbot development in various domains. The proposed work[44] classifies text for a chatbot application in automatic warehouse assistance services using text mining techniques. The Business Process Modeling Notation (BPMN) models are used to connect technological improvements and relationship marketing in chatbot assistance. A two-step process model is used, including hierarchical clustering and Tag Cloud, to identify critical issues customers face. The approach is suitable for automatically creating a combination of chatbot questions and appropriate answers in intelligent systems.

[45]This research focuses on developing a chatbot application for university students to provide educational information. The study uses 1,094 conversations from the Messenger Facebook page of the Faculty of Information Technology during the 2020-2021 session. Data mining and machine learning techniques were used, as well as cross-validation and confusion matrix techniques. The model achieved 88.73% accuracy and an average of 3.97 for application satisfaction. The researchers plan to apply their findings to future academic programs.

1.4 Business Intelligence

Text mining techniques have been employed in business intelligence to extrapolate useful information from enormous amounts of textual data, such as news articles, financial reports, and market research reports.[46] utilize text mining to analyze news articles and predict stock price movements based on sentiment analysis. In [47], the authors draw attention to the growing availability of substantial amounts of textual data in the financial sector and the necessity for effective techniques to extract insightful information from this data and focus on the potential text mining advantages for the financial sector, such as improved decision-making, risk management, and customer satisfaction.

This study [48] examines the use of text mining and business analytics to ambiguous data. There is now ambiguity in outcomes due to the imprecise, inaccurate, and incomplete nature of data in commercial areas. Semantic webs are used by text mining to identify material based on context and meaning, enhancing search and business intelligence outcomes. Counterintelligence and social network analysis are made easier with this method. There isn't much work being done in text data mining, though. the study [48] seeks to investigate emerging trends in business intelligence (BI), including multi-touch, cloud computing, predictive analytics, data visualization, mobile BI, green computing, social networking, and Software-as-a-Service (SaaS).in research [49] uses text mining and latent Dirichlet allocation models to investigate business intelligence applications in the banking sector. It pinpoints credit as the primary application trend in banking, forecasting risk and bolstering acceptance or rejection of credit. The report also emphasizes interest in fraud and bankruptcy forecasting. Relevant publications were used to verify the analysis. In [50], internet abuse in the workplace is increasing, causing productivity losses, resource wasting, security risks, and legal liabilities. Organizations are adopting Internet usage policies, management training, and monitoring. A text mining

approach is proposed for internet abuse detection, promising to complement existing filtering techniques. Experimental results are promising.

This research[51] explores sustainability disclosure frameworks for container shipping companies using a hierarchical unsupervised text-mining method. The framework consists of three primary dimensions: employee training and management, sustainable business management, and sustainable shipping operation. The findings provide insights for container shipping companies, regulators, policymakers, and investors, as well as benchmarking against broader sustainability goals like the United Nations' Sustainable Development Goals.

This study[52] uses text mining to analyze the capabilities of entry-level HR professionals based on job advertisements on HR agency 104's Taiwan website. Python was used to crawl 841 posts, uncovering hidden trends. The results reveal four critical success factors, five clusters, and ten classifications, aiding HR curriculum developers in improving curricula for employment.

This paper [53] provides a comprehensive understanding of the Circular Economy (CE) by analyzing 172 definitions from 2005 to present. The analysis identified 12 topics with 10 keywords, categorized into activities, strategies, aims, and business models. A new, comprehensive definition was proposed, extending the 4R framework and incorporating environmental quality and social harmony considerations. This comprehensive understanding could serve as a foundation for future work and implementation in the CE community.

This study[54] examines energy supply chains in the context of sustainable development using bilateral analysis methodology and performance analysis. It aims to investigate interest in renewable energy supply chains and their impact on sustainability. The analysis provides an overview of current research in this area, offering new possibilities for interpreting and applying management tools. Co-dependency and co-occurrence analysis and text mining provide a foundation for further research.

2. Basic Concepts and Methodologies

Natural language processing (NLP) methods are used in text mining to glean information from unstructured text data. These approaches enhance decision-making procedures and client preferences by analyzing and comprehending massive amounts of textual data. The following are the fundamental components of these approaches:

2.1 Preprocessing

Text preparation techniques are crucial to natural language processing (NLP) operations because they enable the conversion of raw text input into a format more suited for modelling and analysis. The issues posed by unstructured textual data may now be successfully handled by academics and practitioners, thanks to recent references' comprehensive support for various text preparation approaches.

Tokenization, one of the core text preprocessing techniques, involves breaking down the text into smaller units known as tokens [55]. This method aids in word or phrase identification, supporting further analysis like part-of-speech tagging or sentiment analysis [56]. Recent references have presented improved tokenization algorithms that can handle complicated language patterns and domain-specific jargon, enhancing the accuracy and efficiency of NLP models.

Stop-word elimination is another helpful tactic involving irrelevant words that occur frequently in a given context [57]. Recently, references have put forth creative methods for locating and removing stop words that consider the intricacies of different languages and the demands of various domains [58][59]. By eliminating these superfluous words, attention may be brought to more critical information, improving the calibre of the following analyses.

Two further preprocessing techniques that aim to return word inflectional forms to their base or root form are stemming and lemmatization. Lemmatization uses linguistic information to identify a word's basic form, while stemming removes prefixes and suffixes from words [56]. Recent references have included sophisticated stemming and lemmatization algorithms and linguistic resources, allowing for improved handling of irregular word forms and enhancing the precision of downstream NLP tasks [48][53].

Normalization, another crucial technique supported by recent references, involves transforming text data into a consistent format[57]. It includes converting all characters to lowercase, removing punctuation marks or special characters, and appropriately handling numerical expressions or abbreviations [61]. Normalization helps in reducing noise in textual data and ensures consistency across different documents or sources.

Furthermore, recent references have also focused on addressing specific challenges related to text preprocessing techniques. For instance, handling noisy or misspelled text has been addressed through spell-checking and correction techniques. Additionally, dealing with text data in multiple languages has been facilitated by developing language-specific preprocessing techniques, including language detection, transliteration, and translation.

The following procedures are some others that may help improve a text-mining algorithm. Words that have similar grammatical properties, such as noun, verb, or adjective status, can be categorized into the same "part of speech" using this method. Depending on the job at hand, incorporating a step to limit the vocabulary to one or a subset of these parts of speech might be useful as a preprocessing step [57].

2.2 Feature Extraction

For tasks involving machine learning and natural language processing (NLP), text feature extraction is essential. It comprises transforming unprocessed text input into numerical formats for modeling and analysis. The following are some of the primary methods for extracting text features:

1. Bag-of-Words: BOW treats text as a collection of distinct words without consideration to word order or grammar. Every document is represented by a vector that shows whether words are included or not [62][63]
2. Term Frequency-Inverse Document Frequency (TF-IDF): This technique evaluates the importance of a term by looking at its frequency in all documents. Rare words are assigned higher weights and more discriminating [64].
4. Named Entity Recognition (NER): In text data, NER detects and defines named entities like names, places, organizations, etc. It makes it easier to retrieve details about particular textual things[65].
5. Part-of-Speech (P.O.S.) Tagging: POS tagging assigns a grammatical tag to each word in a phrase, such as a noun, verb, adjective, etc. These tags can collect syntactic data as features[66].

Advanced text feature extraction techniques include:

1. Word Embeddings: Word embeddings that capture semantic links by representing words as dense vectors in a multi-dimensional space include Word2Vec, GloVe, and FastText. These embeddings, which may capture contextual and semantic information about words, are learnt from enormous corpora[67].
2. BERT (Bidirectional Encoder Representations from Transformers): BERT is a transformer-based model that analyzes the complete input phrase to determine the contextual embeddings of individual words. Its performance in a range of natural language processing tasks has been impressive.
3. ELMo (Embeddings from Language Models): ELMo generates word embeddings based on deep contextualized word representations learned from the entire input sentence using bidirectional LSTMs.
4. Doc2Vec: This technique extends the concept of word embeddings to entire documents, creating document embeddings that capture the semantic meaning of the entire document.
6. Topic Modeling: Latent Dirichlet Allocation (LDA) and other topic modeling methods detect latent themes in a collection of texts. These topics can express the text's content as features[68], [69].
7. Sentiment Analysis: The characteristics associated with the sentiment or emotion represented in the text are extracted using sentiment analysis. Lexicon-based methods, machine learning models, or deep learning procedures can all be used [70].

2.3 Machine Learning and Deep Learning

A number of significant machine learning and deep learning approaches address text mining. These methods allow a variety of applications, including sentiment analysis, document categorization, topic modeling, and text synthesis, by extracting useful information from textual input. Below are a few of the well-known methods.

Sentiment analysis, which seeks to ascertain the sentiment or opinion expressed in a text, has recently been emphasized. Deep learning models for sentiment analysis have been the subject of recent studies [41]. For instance, [71] suggested a unique deep learning architecture leveraging hierarchical attention networks for fine-grained sentiment analysis. In [72], the authors provide Dependency tree-based Word Embedding (Dt-WE), a new approach for improving the performance of Bidirectional Long Short-Term Memory (Bi-LSTM) models in implicit aspect extraction. Dt-WE generates word embeddings using dependency tree structures, which capture syntactic and semantic links between words. According to experimental data, Dt-WE greatly surpasses previous techniques in accuracy and F1 score.

While [73] highlights the effectiveness of combining machine learning algorithms with augmented data for intent classification tasks in NLP systems, the goal is to determine the intention or purpose behind a user's input. It provides valuable insights for improving the performance of such systems in understanding user intentions accurately.

A method for locating latent themes in a group of texts is called topic modeling. Researchers have been working hard on enhancing topic modeling algorithms and their applications [68], [69]. Regarding subject coherence and interpretability, [74], [75] offered a unique topic modeling methodology based on graph convolutional networks that performed better than conventional techniques.

Named entity recognition (NER) is finding and categorizing named entities in text data, such as names, organizations, places, etc. Contextual embeddings and transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) have been the focus of recent developments in NER.

For instance, [76] suggested a BERT-based NER model that demonstrated state-of-the-art performance on benchmark datasets. Assigning textual documents to specified groups or labels based on their content is another aspect of text categorization. In this field, deep learning models have produced encouraging outcomes. [9] Convolutional neural networks (CNN) and long short-term memory (LSTM) networks were combined to create a hybrid model for text categorization tasks. This model outperformed more conventional techniques in terms of accuracy. Extracting structured data from unstructured text data is another aspect of information extraction. Utilizing deep learning

algorithms for information extraction tasks has been the subject of recent studies. [9] introduced a deep learning-based approach for relation extraction, which outperformed traditional rule-based methods in terms of precision and recall.

Text Generation: Text generation creates cohesive, insightful writing in response to conditions or suggestions. Text creation challenges have traditionally been handled through methods like Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). Recent studies have concentrated on applying methods like reinforcement learning to increase the quality and variety of produced text [77].

Finally, text summarization aims to generate concise summaries of longer texts. Recent studies have explored transformer-based models for text summarization tasks[6]. For instance, [76] proposed a transformer-based model called BART (Bidirectional and Auto-Regressive Transformers) for abstractive text summarization, which achieved state-of-the-art performance on multiple benchmark datasets.

3. Emerging Trends and Opportunities

Text mining research has witnessed significant advancements and emerging trends since 2022. With the exponential growth of digital data, there is a pressing need to extract valuable insights from unstructured text sources. Integrating deep learning techniques like RNNs and transformers into text mining research is a current topic to enhance the precision and performance of natural language processing applications. Named entity recognition, sentiment analysis, and question answering are just some of the text-mining applications that BERT has transformed [78]. Another developing pattern is adapting text mining methods to accommodate several languages and cultures. To overcome the language barrier in text analysis, researchers have created new methods using cross-lingual embeddings and transfer learning [79]. In addition, there is a growing emphasis on moral issues in text mining research, mainly regarding confidentiality and bias mitigation [80]. Recent studies have proposed methods for ensuring fairness and transparency in automated decision-making systems based on text-mining algorithms. Overall, these emerging trends present exciting opportunities for advancing the field of text mining research and addressing real-world challenges in various domains [59][57][81]. Chatbots are another recent trend. A chatbot is a computer software that simulates conversation with human users, typically through the internet[82].

The study uses a survey method to collect data from marketing professionals with chatbot implementation experience. Lately, GPT's (Generative Pre-trained Transformer) revolution has sparked a revolution in natural language processing and artificial intelligence. GPT, a deep learning model created by OpenAi, employs a transformer architecture to produce human-like language in response to provided cues. Due to its capability to produce replies that are both logical and contextually appropriate, making it practically impossible to tell them apart from human-generated content, it has attracted considerable attention and popularity. Numerous applications, including chatbots, language translation, content creation, and creative writing, have extensively used GPT [83]. Its impact on industries like journalism, customer service, and marketing has been profound, as it can automate tasks that previously required human intervention. However, concerns have also been raised regarding the ethical implications of using GPT for generating fake news or spreading misinformation. Despite these concerns, the GPTrevolution continues to evolve rapidly with ongoing research and advancements in A.I. technology[77][78].

4. A.I. Integrations

Text mining has improved its capacity for multimodal analysis by combining with other artificial intelligence (AI) tools like speech or image recognition. Combining these technologies enables academics and practitioners to concurrently evaluate and comprehend data from many sources, producing more thorough and precise conclusions. [86], [87].

Social media analysis is one area where text mining and image recognition have been combined. The extraction of useful information from images has grown essential [88] due to the exponential development of visual materials published on platforms like Instagram or Pinterest. Researchers have created methods that examine visual content and supporting written descriptions by fusing text-mining techniques with image recognition algorithms. This connection makes it possible to analyze user sentiment, forecast trends based on visual clues, and gain a more excellent knowledge of user preferences[86], [87].

The combination of text mining and speech recognition has introduced novel opportunities in domains such as customer service and healthcare. Through the examination of transcriptions or recordings of customer interactions or medical consultations[89], artificial intelligence (AI) systems have the capability to derive significant insights pertaining to customer satisfaction levels or the detection of probable health concerns. Multimodal analysis empowers firms to enhance their services and deliver tailored experiences by comprehensively understanding customer wants. Current research has emphasized the development of sophisticated models that integrate text mining with other forms of AI technologies to enable multimodal analysis. Researchers have put forward deep learning architectures that combine natural language processing techniques and computer vision algorithms to accomplish tasks such as picture captioning or sentiment analysis by utilizing textual and visual data[90][91].

Knowledge graphs and ontologies are powerful tools for improving textual data understanding. With the use of ontology-based text-mining algorithms, ontologies provide a methodical framework for arranging and accessing

knowledge, making it easier to extract structured data from unstructured text [84] [85]. The use of ontologies to improve text understanding and speed up the creation of complex applications like recommendation and semantic search systems is examined in this article. Furthermore, it delves into the use of knowledge graphs in text comprehension assignments including relation extraction, entity linking, and question answering. It goes over several methods for creating knowledge graphs from textual data and emphasizes how they may help with text comprehension [86] [87]

5. Challenges

When using text mining techniques, there are a number of obstacles to overcome. Several of the major obstacles consist of:

1. Integrating text data from many sources and formats while upholding standards and compatibility to provide easy analysis and insights is known as data integration and interoperability.
2. Clinical Context Understanding: Creating strategies for interpreting medical terminology, context-specific acronyms, and the subtleties of patient records in order to understand the clinical context of medical literature.
3. Ethical and Privacy Concerns: Addressing the ethical and privacy issues associated with analyzing sensitive medical data, including health information and patient records, while guaranteeing adherence to laws such as the United States' HIPAA (Health Insurance Portability and Accountability Act).
4. Real-time Analysis and Decision Assistance: Allowing real-time text mining to respond promptly to social network trends and marketing dynamics, or to provide rapid decision assistance in healthcare situations.
5. Multimodal Data Analysis: Combining and evaluating textual data with additional data formats, such as audio, video, and pictures, to gain thorough understanding and support decision-making.
6. Bias and Fairness: Minimizing biases and ensuring fairness, particularly in delicate fields like social network analysis and healthcare, to avoid discrimination and unfair results.
7. Semi-Structured Data: Creating methods for managing semi-structured data, such as data inserted into forms, structured documents, or templates.
8. Explainability and Trust: Improving the explainability and interpretability of text mining models and findings will help stakeholders, especially those involved in marketing and medical decision-making, feel more confident.
9. Dynamic Language and Slang: Modifying text-mining techniques to account for language fluidity, which includes colloquialisms, slangs, and quickly changing terminology in marketing communications and social network material.
10. Cross-disciplinary Collaboration: Facilitating collaboration between data scientists, domain experts, and stakeholders from many domains is facilitated to guarantee significant insights and useful results from text mining analyses.

6. Limitations

Text mining techniques have garnered considerable interest in recent times due to their capacity to extract meaningful insights from extensive quantities of unstructured textual data. Nevertheless, the utilization of these methodologies may be subject to specific constraints when applied in diverse fields and businesses.

One constraint that must be acknowledged is the inherent limitations with respect to the quality and dependability of the textual material under consideration. The efficacy and comprehensiveness of text mining are strongly contingent upon the precision and entirety of the input data. If the text data is noisy, contains errors, or lacks relevant information, it can negatively impact the effectiveness of text mining techniques. Additionally, text mining may struggle with understanding sarcasm, irony, or other forms of figurative language that are commonly used in textual communication [96].

Another limitation is related to privacy and ethical concerns. Text mining often involves analyzing personal or sensitive information from sources such as social media posts or customer reviews. Ensuring proper anonymization and safeguarding individuals' privacy becomes crucial in such cases [97], [98]. Moreover, ethical considerations arise when using text mining techniques for sentiment analysis or opinion mining, as misinterpretation or manipulation of textual data can lead to biased results [96].

Furthermore, domain-specific challenges can hinder the application of text mining techniques across different industries. Each industry has its own unique vocabulary and context-specific nuances that need to be considered during analysis. For example, medical texts may contain complex terminology that requires specialized knowledge for accurate interpretation [99], [100]. Similarly, legal texts may involve intricate legal jargon that necessitates domain expertise for effective extraction of relevant information.

7. Discussion

Text mining techniques have received a lot of attention in recent years due to the tremendous growth in the availability of text data in various industries such as social media, healthcare, finance, and e-commerce. An important

step in text mining is feature extraction, which involves transforming the raw text into a statistical representation that can be used for further analysis.

Text mining strategies have evolved over the years to cope with the challenges posed by way of the growing volume and complexity of textual information. Initially, traditional strategies including methods primarily based on keywords were used for characteristic extraction. These techniques relied on predefined lists of key phrases or terms to discover applicable facts from the textual content. However, they often suffered from confined coverage and lacked the ability to seize semantic relationships among phrases.

With the development of natural language processing (NLP) and machine learning (ML), advanced methods have emerged for the production of text mining features. Another popular method is word embedding, which represents words as dense vectors in a high-dimensional space. Word embedding captures the semantic relationships between words by considering their meaning within a large amount of textual information. Methods such as Word2Vec are now widely accepted for word input.

Another emerging tendency in feature extraction is the utilization of advanced deep learning architectures, such as RNN and CNN. Recurrent Neural Networks (RNNs) have demonstrated notable efficacy in capturing sequential relationships in textual input, rendering them well-suited for applications such as sentiment analysis and named entity recognition. Contrarily, CNNs demonstrate exceptional proficiency in capturing intricate local patterns, proving their efficacy in various applications such as document classification and topic modelling.

However, text-mining methodologies are often faced with several challenges. One notable challenge pertains to the effective handling and organization of voluminous and disorganized textual material. Textual data sometimes includes spelling problems, abbreviations, colloquial language, and grammatical errors, all of which might potentially affect the effectiveness of text-mining algorithms. Preprocessing techniques, including tokenization, stemming, and lemmatization, are commonly utilized to address the aforementioned challenges. Another challenge that emerges is the lack of well annotated training data for approaches that utilize supervised learning. The task of assigning precise labels to extensive amounts of textual content can be arduous and expensive. Active learning and transfer learning are two methods that have been studied by researchers as potential solutions to the problem of working with insufficient labeled data. It is crucial to recognize the considerable limits of text mining algorithms when attempting to capture the contextual and semantic characteristics of textual data. While word embeddings may be limited in their ability to capture sophisticated semantic links, they still provide a useful starting point. The difficult task of detecting irony, sarcasm, or minute emotional inconsistencies in textual data is still far from being fully handled by text mining algorithms.

Recent years have seen a substantial advancement in text mining techniques. Deep learning models and word embeddings are two examples of the more sophisticated techniques being used. The aforementioned techniques have shown promising outcomes when it comes to feature extraction from text data. However, because of problems like noisy data, unlabeled training data, and inaccurately capturing complicated semantics, the field still has to be addressed. To address these problems, more investigation and advancement are needed.

Within the social media domain, text mining is essential for making sense of and applying to the enormous volumes of user-generated material. Organizations may learn a great deal about the attitudes, tastes, and actions of their customers by examining social media postings, comments, and correspondence. A common text mining method used by corporations to evaluate the success of their social media efforts, discover new trends, and determine public opinion is sentiment analysis. Strategies for product creation, consumer interaction, and strategic decision-making may all benefit from these insights. Identifying influencers, tracking brand reputation, and spotting new problems are further uses for social media text mining.

Within the field of business and marketing, text mining makes it easier to analyze competition data, market trends, and client feedback. Businesses may obtain a thorough grasp of client preferences, problems, and new requests by gathering and evaluating data from sources including industry reports, customer evaluations, and survey replies. This may support the creation of focused marketing efforts, the improvement of product offers, and the enhancement of the customer experience. Moreover, through the extraction of insights from publicly accessible textual data, text mining helps firms do competitive research, which enables them to keep current with market dynamics and modify their strategy appropriately.

In order to enhance patient safety, research results, and healthcare delivery, text mining is crucial for the analysis of clinical notes, research papers, and patient data. Within unstructured clinical data, text mining can assist in finding pertinent medical entities, connections, and trends using methods like Named Entity Recognition and information extraction. This supports pharmacovigilance, adverse event identification, and clinical decision assistance. Additionally, text mining advances biomedical research by making it possible to extract useful data from a vast number of clinical trials and scholarly publications. It supports evidence-based medicine, helps discover new therapeutic targets, and comprehends disease mechanisms.

All things considered; text mining is an effective method for turning unstructured text data into insights that can be used to a variety of different fields. Organizations may use text mining to promote innovation, enhance decision-making, and achieve a competitive edge in their respective sectors by utilizing advanced natural language processing

techniques, machine learning algorithms, and domain-specific expertise. To ensure the ethical and efficient application of text mining in a variety of fields, it is imperative to address data privacy, data quality, and ethical issues.

8. Conclusion

To sum up, text mining has developed into a critical tool that makes it possible to extract information and important insights from a vast amount of unstructured textual data. Its uses cut across a wide range of applied computer applications areas, thereby advancing a multitude of fields. Text mining has greatly enhanced language creation and interpretation in the field of natural language processing, leading to the development of intelligent chatbots, sentiment analysis, and language translation services. Furthermore, text mining is an essential tool in the business and finance arena that can be used to analyze market trends, consumer feedback, and financial reports. This allows for the analysis of data to drive decision-making processes and improve corporate strategies. It is impossible to overestimate the significance of text mining in the healthcare industry as it has made it easier to analyze patient data, medical literature, and medication discovery, which has improved clinical results and advanced public health. Furthermore, in cybersecurity, text mining has shown promise in identifying and mitigating potential security threats through the analysis of large volumes of textual data. Legal practitioners have also benefited from text mining applications, aiding in contract analysis, e-discovery, and legal research. The complex landscape of social media is another domain where text mining has proven invaluable, enabling sentiment analysis, trend detection, and opinion mining. As text mining techniques continue to advance, their transformative impact on various applied computer applications domains is expected to grow, opening up new frontiers for innovation and discovery.

Funding

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

None

9. References

- [1] K. Thakur and V. Kumar, "Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools," *New Rev. Acad. Librariansh.*, vol. 28, no. 3, pp. 279–302, 2022.
- [2] G. Smith, "Data mining fool's gold," *J. Inf. Technol.*, vol. 35, no. 3, pp. 182–194, May 2020.
- [3] J. Zeng *et al.*, "Operationalization of Next-Generation Sequencing and Decision Support for Precision Oncology," *JCO Clin. Cancer Informatics*, no. 3, pp. 1–12, 2019.
- [4] J.-Y. Huang and J.-H. Liu, "Using social media mining technology to improve stock price forecast accuracy," *J. Forecast.*, vol. 39, no. 1, pp. 104–116, 2020.
- [5] Y. Zhao, X. Xu, and M. Wang, "Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews," *Int. J. Hosp. Manag.*, vol. 76, pp. 111–121, 2019.
- [6] R. A. Albeer, H. F. Al-Shahad, H. J. Aleqabie, and N. D. Al-Shakarchy, "Automatic summarization of YouTube video transcription text using term frequency-inverse document frequency," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 3, pp. 1512–1519, 2022.
- [7] L. T. Wu, J. R. Lin, S. Leng, J. L. Li, and Z. Z. Hu, "Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web," *Autom. Constr.*, vol. 135, no. January, p. 104108, 2022.
- [8] S. Tejasree and S. Naseera, "Improved Clustering Technique Using Metadata for Text Mining," in *Innovations in Computer Science and Engineering: Proceedings of the Sixth ICICSE 2018*, 2019, pp. 243–250.
- [9] X. She and D. Zhang, "Text Classification Based on Hybrid CNN-LSTM Hybrid Model," *Proc. - 2018 11th Int. Symp. Comput. Intell. Des. Isc. 2018*, vol. 2, pp. 185–189, 2018.
- [10] "No Title."
- [11] D. A. Naik, S. Mythreyan, and S. Seema, "Relevance Feature Discovery in Text Mining Using NLP," in *2022 3rd International Conference for Emerging Technology (INCET)*, 2022, pp. 1–6.
- [12] Y. Ding, J. Ma, and X. Luo, "Applications of natural language processing in construction," *Autom. Constr.*, vol. 136, p. 104169, 2022.
- [13] A. Alomari, N. Idris, A. Q. M. Sabri, and I. Alsmadi, "Deep reinforcement and transfer learning for abstractive text summarization: A review," *Comput. Speech Lang.*, vol. 71, p. 101276, 2022.
- [14] D. Suleiman and A. Awajan, "Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges," *Math. Probl. Eng.*, vol. 2020, p. 9365340, 2020.

- [15] Y. Kim, S. Bang, J. Sohn, and H. Kim, "Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers," *Autom. Constr.*, vol. 134, p. 104061, 2022.
- [16] J. Lin, "A proposed conceptual framework for a representational approach to information retrieval," *ACM SIGIR Forum*, vol. 55, no. 2, pp. 1–29, 2021.
- [17] N. Eligüzel, C. Cetinkaya, and T. Dereli, "A novel approach for text categorization by applying hybrid genetic bat algorithm through feature extraction and feature selection methods," *Expert Syst. Appl.*, vol. 202, p. 117433, 2022.
- [18] Q. Peng, D. Weir, J. Weeds, and Y. Chai, "Predicate-argument based bi-encoder for paraphrase identification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5579–5589.
- [19] C. Perfetti and A. Helder, "Progress in reading science: Word identification, comprehension, and universal perspectives," *Sci. Read. A Handb.*, pp. 5–35, 2022.
- [20] F. Liu, C. Weng, and H. Yu, "Advancing clinical research through natural language processing on electronic health records: traditional machine learning meets deep learning," *Clin. Res. Informatics*, pp. 357–378, 2019.
- [21] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "Am i no good? towards detecting perceived burdensomeness and thwarted belongingness from suicide notes," *arXiv Prepr. arXiv2206.06141*, 2022.
- [22] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, pp. 214–226, 2020.
- [23] S. Pal, B. Biswas, R. Gupta, A. Kumar, and S. Gupta, "Exploring the factors that affect user experience in mobile-health applications: A text-mining and machine-learning approach," *J. Bus. Res.*, vol. 156, no. December 2022, p. 113484, 2023.
- [24] Z. Zhang, T. Jasaitis, R. Freeman, R. Alfrjani, A. Funk, and R. Court, "Mining Healthcare Procurement Data Using Text Mining and Natural Language Processing -- Reflection From An Industrial Project. (arXiv:2301.03458v1 [cs.CL])," *arXiv Comput. Sci.*, 2022.
- [25] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: A review," *J. Healthc. Eng.*, vol. 2018, 2018.
- [26] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," *Proc. ACM Symp. Appl. Comput.*, vol. 1, pp. 235–239, 2006.
- [27] P. Thompson *et al.*, "Text mining the history of medicine," *PLoS One*, vol. 11, no. 1, pp. 1–33, 2016.
- [28] C. Sciences, "Text mining of clinical outcomes for medical research : how accurate should it be ?," 2022.
- [29] A. Al-Aiad and T. El-shqeirat, "Text mining in radiology reports (Methodologies and algorithms), and how it affects on workflow and supports decision making in clinical practice (Systematic review)," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020, pp. 283–287.
- [30] N. Chintalapudi *et al.*, "LASSO Regression Modeling on Prediction of Medical Terms among Seafarers' Health Documents Using Tidy Text Mining," *Bioengineering*, vol. 9, no. 3, pp. 1–14, 2022.
- [31] A. W. K. Yeung *et al.*, "Research on Digital Technology Use in Cardiology: Bibliometric Analysis," *J Med Internet Res*, vol. 24, no. 5, p. e36086, May 2022.
- [32] D. D. Martinelli, "Evolution of Alzheimer's disease research from a health-tech perspective: Insights from text mining," *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 2, p. 100089, 2022.
- [33] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.
- [34] M. AbdelFattah, D. Galal, N. Hassan, D. Elzanfaly, and G. Tallent, "A Sentiment Analysis Tool for Determining the Promotional Success of Fashion Images on Instagram.," *Int. J. Interact. Mob. Technol.*, vol. 11, no. 2, pp. 66–73, 2017.
- [35] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment analysis for fake news detection," *Electron.*, vol. 10, no. 11, 2021.
- [36] H. Xu and Y. Lv, "Mining and Application of Tourism Online Review Text Based on Natural Language Processing and Text Classification Technology," *Wirel. Commun. Mob. Comput.*, vol. 2022, p. 9905114, 2022.
- [37] M. Alzate, M. Arce-Urriza, and J. Cebollada, "Mining the text of online consumer reviews to analyze brand image and brand positioning," *J. Retail. Consum. Serv.*, vol. 67, p. 102989, 2022.
- [38] Y.-J. Kim and H.-S. Kim, "The Impact of Hotel Customer Experience on Customer Satisfaction through Online Reviews," *Sustainability*, vol. 14, no. 2, 2022.
- [39] J. Y. Park, E. Mistur, D. Kim, Y. Mo, and R. Hoefer, "Toward human-centric urban infrastructure: Text mining for social media data to identify the public perception of COVID-19 policy in transportation hubs.," *Sustain. cities Soc.*, vol. 76, p. 103524, Jan. 2022.
- [40] X. Yu, Y. Liu, X. Huang, and A. An, "Mining online reviews for predicting sales performance: A case study in the movie domain," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 720–734, 2010.
- [41] H. J. Aleqabie, M. S. Safoq, I. R. Shareef, R. Alsabah, and E. H. Abd, "Sentiment analysis for movie reviews using embedding words with semantic orientation," *AIP Conf. Proc.*, vol. 2290, no. December, 2020.

- [42] F. Janin Jalal and A. Al-Haj Hussein, "Impact of Customer Relationship Management on Customer Satisfaction in Private Sector Organization," pp. 1–13, 2018.
- [43] B. Galitsky, "Chatbots for CRM and Dialogue Management," in *Artificial Intelligence for Customer Relationship Management: Solving Customer Problems*, Cham: Springer International Publishing, 2021, pp. 1–61.
- [44] A. Massaro, N. Magaletti, G. Cosoli, V. Giardinelli, and A. Leogrande, "Text Mining Approaches Oriented on Customer Care Efficiency," *SSRN Electron. J.*, pp. 0–28, 2022.
- [45] P. Nasa-Ngium, W. S. Nuankaew, and P. Nuankaew, "Analyzing and Tracking Student Educational Program Interests on Social Media with Chatbots Platform and Text Analytics," *Int. J. Interact. Mob. Technol.*, vol. 17, no. 5, pp. 4–21, 2023.
- [46] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, "Mine your own business: Market-structure surveillance through text mining," *Mark. Sci.*, vol. 31, no. 3, pp. 521–543, 2012.
- [47] M. Pejić Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," *Sustainability*, vol. 11, no. 5, p. 1277, 2019.
- [48] P. Gupta and B. Narang, "Role of text mining in business intelligence," *Gian Jyoti E-Journal*, vol. 1, no. 2, pp. 1–13, 2012.
- [49] E. Systems and E. Systems, "Business Intelligence in Banking : a Literature Analysis from 2002 to 2013 using Text Mining and Latent Dirichlet Allocation," no. 351, 2019.
- [50] C.-H. Chou, A. P. Sinha, and H. Zhao, "A text mining approach to Internet abuse detection," *Inf. Syst. E-bus. Manag.*, vol. 6, no. 4, pp. 419–439, 2008.
- [51] Y. Zhou, X. Wang, and K. F. Yuen, "Sustainability disclosure for container shipping: A text-mining approach," *Transp. Policy*, vol. 110, pp. 465–477, 2021.
- [52] C.-H. Chung and L.-J. Chen, "Text mining for human resources competencies: Taiwan example," *Eur. J. Train. Dev.*, vol. 45, no. 6/7, pp. 588–602, Jan. 2021.
- [53] M. Alizadeh, A. Kashef, Y. Wang, J. Wang, G. E. Okudan Kremer, and J. Ma, "Circular economy conceptualization using text mining analysis," *Sustain. Prod. Consum.*, vol. 35, pp. 643–654, 2023.
- [54] B. Tundys and T. Wiśniewski, "Renewable Energy Supply Chains—Text Mining and Co-Occurrence Analysis in the Context of the Sustainability," *Energies*, vol. 16, no. 12, 2023.
- [55] M. Işık and H. Dağ, "The impact of text preprocessing on the prediction of review ratings," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 28, no. 3, pp. 1405–1421, 2020.
- [56] V. Gurusamy and Kannan S, "Preprocessing Techniques for Text Mining," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2014.
- [57] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, " Survey on Text Classification Algorithms: From Text to Predictions," *Inf.*, vol. 13, no. 2, pp. 1–39, 2022.
- [58] P. Y. Shotorbani, "Text Mining Techniques for Analyzing Unstructured," Texas State University, 2016.
- [59] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, 2022.
- [60] M. Saad, "The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification," The Islamic University-Gaza, 2010.
- [61] S. Pradha, M. N. Halgamuge, and N. Tran Quoc Vinh, "Effective text data preprocessing technique for sentiment analysis in social media data," *Proc. 2019 11th Int. Conf. Knowl. Syst. Eng. KSE 2019*, pp. 1–8, 2019.
- [62] H. K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation," *Neurocomputing*, vol. 266, pp. 336–352, 2017.
- [63] Y. Zhang, M. Chen, and L. Liu, "A review on text mining," in *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2015, pp. 681–685.
- [64] B. Parlak, "A novel feature and class-based globalization technique for text classification," *Multimed. Tools Appl.*, pp. 1–26, 2023.
- [65] O. Bridal, "Named-entity recognition with BERT for anonymization of medical records." 2021.
- [66] Y. B. Kim, B. Snyder, and R. Sarikaya, "Part-of-speech taggers for low-resource languages using CCA features," *Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process.*, no. September, pp. 1292–1302, 2015.
- [67] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv Prepr. arXiv1810.04805*, 2018.
- [68] G. A. Al-Sultany and H. J. Aleqabie, "Enriching Tweets for Topic Modeling via Linking to the Wikipedia," *Int. J. Eng. Technol.*, vol. 7, no. 19, pp. 144–150, 2018.
- [69] G. A. Al-Sultany and H. J. Aleqabie, "Events Tagging in Twitter Using Twitter Latent Dirichlet Allocation," *Int. J. Eng. Technol.*, vol. 7, no. 4.19, pp. 884–888, 2018.
- [70] A. Bello, S. C. Ng, and M. F. Leung, "A BERT Framework to Sentiment Analysis of Tweets," *Sensors*, vol. 23,

- no. 1, 2023.
- [71] S. N. inglin Shao and Beijing, “Fine-Grained Sentiment Analysis Based on Hierarchical Attention Networks,” *Comput. Sci. Appl.*, vol. 09, no. 11, pp. 2143–2153, 2019.
- [72] O. M. Al-Janabi, M. K. Ibrahim, A. Kanaan-Jebna, O. M. Alyasiri, and H. J. Aleqabie, “An improved Bi-LSTM performance using Dt-WE for implicit aspect extraction,” in *2022 International Conference on Data Science and Intelligent Computing, ICDSIC 2022*, 2022, pp. 14 – 19.
- [73] A. T. Al-Tuama and D. A. Nasrawi, “Intent Classification Using Machine Learning Algorithms and Augmented Data,” *2022 Int. Conf. Data Sci. Intell. Comput. ICDSIC 2022*, no. Icdsic, pp. 234–239, 2022.
- [74] J. Zhang *et al.*, “Graph convolutional network-strengthened topic modeling for scientific papers,” *Proc. - 2021 IEEE Int. Conf. Smart Data Serv. SMDs 2021*, pp. 24–32, 2021.
- [75] R. Egger and J. Yu, “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts,” *Front. Sociol.*, vol. 7, no. May, pp. 1–16, 2022.
- [76] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.
- [77] P. Laban, L. Dai, L. Bandarkar, and M. A. Hearst, “Can Transformer Models Measure Coherence in Text? Rethinking the Shuffle Test,” *ACL-IJCNLP 2021 - 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, vol. 2, pp. 1058–1064, 2021.
- [78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “{BERT}: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
- [79] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual BERT?,” *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 4996–5001, 2020.
- [80] T. Bolukbasi, K. W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings,” *Adv. Neural Inf. Process. Syst.*, pp. 4356–4364, 2016.
- [81] M. Lamba and M. Madhusudhan, *Text Mining for Information Professionals*. 2022.
- [82] A. T. Al-Tuama and D. A. Nasrawi, “A Survey on the Impact of Chatbots on Marketing Activities,” *2022 13th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2022*, no. October, 2022.
- [83] C. Zhang *et al.*, “A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need?,” *arXiv Prepr. arXiv2303.11717*, 2023.
- [84] A. Radford *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [85] X. Liu *et al.*, “Large Language Models are Few-Shot Health Learners,” *arXiv Prepr. arXiv2305.15525*, 2023.
- [86] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. M. Sadeeq, and S. Zeebaree, “Multimodal emotion recognition using deep learning,” *J. Appl. Sci. Technol. Trends*, vol. 2, no. 02, pp. 52–58, 2021.
- [87] W. Zhai, X. Bai, Y. Shi, Y. Han, Z.-R. Peng, and C. Gu, “Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs,” *Comput. Environ. Urban Syst.*, vol. 74, pp. 1–12, 2019.
- [88] D. J. Mohammed and H. J. Aleqabie, “The Enrichment Of MVSA Twitter Data Via Caption-Generated Label Using Sentiment Analysis,” in *2022 Iraqi International Conference on Communication and Information Technologies (ICCIT)*, 2022, pp. 322–327.
- [89] J. K. Buser, Y.-J. Cheng, and R. A. McLaughlin, “Thematic Analysis,” *Reimagining Res. Engag. Data, Res. Progr. Eval. Soc. Justice Couns.*, p. 153, 2023.
- [90] D. J. Mohammed and H. J. Aleqabie, “Context-Based Visual Sentiment Analysis for Social Media Data,” in *2022 International Conference on Artificial Intelligence of Things (ICAIoT)*, 2022, pp. 1–5.
- [91] L. Tan *et al.*, “Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space--air--ground integrated intelligent transportation system,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2830–2842, 2021.
- [92] B. Müller, L. J. Castro, and D. Rebolz-Schuhmann, “Ontology-based identification and prioritization of candidate drugs for epilepsy from literature,” *J. Biomed. Semantics*, vol. 13, no. 1, pp. 1–18, 2022.
- [93] G.-K. J. Li, C. V Trappey, A. J. C. Trappey, and A. A. S. Li, “Ontology-based knowledge representation and semantic topic modeling for intelligent trademark legal precedent research,” *World Pat. Inf.*, vol. 68, p. 102098, 2022.
- [94] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, “Knowledge Graphs: Opportunities and Challenges,” *Artif. Intell. Rev.*, 2023.
- [95] P. A. Bonatti, S. Decker, A. Polleres, and V. Presutti, “Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web,” *Rep. from Dagstuhl Semin.*, vol. 8, no. 09, p. 18371, 2018.
- [96] M. Wankhade, A. C. S. Rao, and C. Kulkarni, “A survey on sentiment analysis methods, applications, and challenges,” *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, 2022.

- [97] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *Springerplus*, vol. 4, no. 1, pp. 1–36, 2015.
- [98] M. Kayaalp, "Patient privacy in the era of big data," *Balkan Med. J.*, vol. 35, no. 1, pp. 8–17, 2018.
- [99] N. Jindal and B. Liu, "Analyzing and detecting review spam," in *Seventh IEEE international conference on data mining (ICDM 2007)*, 2007, pp. 547–552.
- [100] M. Salehan and D. J. Kim, "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics," *Decis. Support Syst.*, vol. 81, pp. 30–40, 2016.
- [1] K. Thakur and V. Kumar, "Application of Text Mining Techniques on Scholarly Research Articles: Methods and Tools," *New Rev. Acad. Librariansh.*, vol. 28, no. 3, pp. 279–302, 2022.
- [2] G. Smith, "Data mining fool's gold," *J. Inf. Technol.*, vol. 35, no. 3, pp. 182–194, May 2020.
- [3] J. Zeng *et al.*, "Operationalization of Next-Generation Sequencing and Decision Support for Precision Oncology," *JCO Clin. Cancer Informatics*, no. 3, pp. 1–12, 2019.
- [4] J.-Y. Huang and J.-H. Liu, "Using social media mining technology to improve stock price forecast accuracy," *J. Forecast.*, vol. 39, no. 1, pp. 104–116, 2020.
- [5] Y. Zhao, X. Xu, and M. Wang, "Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews," *Int. J. Hosp. Manag.*, vol. 76, pp. 111–121, 2019.
- [6] R. A. Albeer, H. F. Al-Shahad, H. J. Aleqabie, and N. D. Al-Shakarchy, "Automatic summarization of YouTube video transcription text using term frequency-inverse document frequency," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 26, no. 3, pp. 1512–1519, 2022.
- [7] L. T. Wu, J. R. Lin, S. Leng, J. L. Li, and Z. Z. Hu, "Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web," *Autom. Constr.*, vol. 135, no. January, p. 104108, 2022.
- [8] S. Tejasree and S. Naseera, "Improved Clustering Technique Using Metadata for Text Mining," in *Innovations in Computer Science and Engineering: Proceedings of the Sixth ICICSE 2018*, 2019, pp. 243–250.
- [9] X. She and D. Zhang, "Text Classification Based on Hybrid CNN-LSTM Hybrid Model," *Proc. - 2018 11th Int. Symp. Comput. Intell. Des. Isc. 2018*, vol. 2, pp. 185–189, 2018.
- [10] "No Title."
- [11] D. A. Naik, S. Mythreyan, and S. Seema, "Relevance Feature Discovery in Text Mining Using NLP," in *2022 3rd International Conference for Emerging Technology (INCET)*, 2022, pp. 1–6.
- [12] Y. Ding, J. Ma, and X. Luo, "Applications of natural language processing in construction," *Autom. Constr.*, vol. 136, p. 104169, 2022.
- [13] A. Alomari, N. Idris, A. Q. M. Sabri, and I. Alsmadi, "Deep reinforcement and transfer learning for abstractive text summarization: A review," *Comput. Speech Lang.*, vol. 71, p. 101276, 2022.
- [14] D. Suleiman and A. Awajan, "Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges," *Math. Probl. Eng.*, vol. 2020, p. 9365340, 2020.
- [15] Y. Kim, S. Bang, J. Sohn, and H. Kim, "Question answering method for infrastructure damage information retrieval from textual data using bidirectional encoder representations from transformers," *Autom. Constr.*, vol. 134, p. 104061, 2022.
- [16] J. Lin, "A proposed conceptual framework for a representational approach to information retrieval," *ACM SIGIR Forum*, vol. 55, no. 2, pp. 1–29, 2021.
- [17] N. Eligüzel, C. Cetinkaya, and T. Dereli, "A novel approach for text categorization by applying hybrid genetic bat algorithm through feature extraction and feature selection methods," *Expert Syst. Appl.*, vol. 202, p. 117433, 2022.
- [18] Q. Peng, D. Weir, J. Weeds, and Y. Chai, "Predicate-argument based bi-encoder for paraphrase identification," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5579–5589.
- [19] C. Perfetti and A. Helder, "Progress in reading science: Word identification, comprehension, and universal perspectives," *Sci. Read. A Handb.*, pp. 5–35, 2022.
- [20] F. Liu, C. Weng, and H. Yu, "Advancing clinical research through natural language processing on electronic health records: traditional machine learning meets deep learning," *Clin. Res. Informatics*, pp. 357–378, 2019.
- [21] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "Am i no good? towards detecting perceived burdensomeness and thwarted belongingness from suicide notes," *arXiv Prepr. arXiv2206.06141*, 2022.
- [22] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, pp. 214–226, 2020.
- [23] S. Pal, B. Biswas, R. Gupta, A. Kumar, and S. Gupta, "Exploring the factors that affect user experience in mobile-health applications: A text-mining and machine-learning approach," *J. Bus. Res.*, vol. 156, no. December 2022, p. 113484, 2023.
- [24] Z. Zhang, T. Jasaitis, R. Freeman, R. Alfrjani, A. Funk, and R. Court, "Mining Healthcare Procurement Data Using Text Mining and Natural Language Processing -- Reflection From An Industrial Project. (arXiv:2301.03458v1 [cs.CL])," *arXiv Comput. Sci.*, 2022.
- [25] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on

- electronic medical records: A review,” *J. Healthc. Eng.*, vol. 2018, 2018.
- [26] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, “Approaches to text mining for clinical medical records,” *Proc. ACM Symp. Appl. Comput.*, vol. 1, pp. 235–239, 2006.
- [27] P. Thompson *et al.*, “Text mining the history of medicine,” *PLoS One*, vol. 11, no. 1, pp. 1–33, 2016.
- [28] C. Sciences, “Text mining of clinical outcomes for medical research : how accurate should it be ?,” 2022.
- [29] A. Al-Aiad and T. El-shqeir, “Text mining in radiology reports (Methodologies and algorithms), and how it affects on workflow and supports decision making in clinical practice (Systematic review),” in *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020, pp. 283–287.
- [30] N. Chintalapudi *et al.*, “LASSO Regression Modeling on Prediction of Medical Terms among Seafarers’ Health Documents Using Tidy Text Mining,” *Bioengineering*, vol. 9, no. 3, pp. 1–14, 2022.
- [31] A. W. K. Yeung *et al.*, “Research on Digital Technology Use in Cardiology: Bibliometric Analysis,” *J Med Internet Res*, vol. 24, no. 5, p. e36086, May 2022.
- [32] D. D. Martinelli, “Evolution of Alzheimer’s disease research from a health-tech perspective: Insights from text mining,” *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 2, p. 100089, 2022.
- [33] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.
- [34] M. AbdelFattah, D. Galal, N. Hassan, D. Elzanfaly, and G. Tallent, “A Sentiment Analysis Tool for Determining the Promotional Success of Fashion Images on Instagram,” *Int. J. Interact. Mob. Technol.*, vol. 11, no. 2, pp. 66–73, 2017.
- [35] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, “Sentiment analysis for fake news detection,” *Electron.*, vol. 10, no. 11, 2021.
- [36] H. Xu and Y. Lv, “Mining and Application of Tourism Online Review Text Based on Natural Language Processing and Text Classification Technology,” *Wirel. Commun. Mob. Comput.*, vol. 2022, p. 9905114, 2022.
- [37] M. Alzate, M. Arce-Urriza, and J. Cebollada, “Mining the text of online consumer reviews to analyze brand image and brand positioning,” *J. Retail. Consum. Serv.*, vol. 67, p. 102989, 2022.
- [38] Y.-J. Kim and H.-S. Kim, “The Impact of Hotel Customer Experience on Customer Satisfaction through Online Reviews,” *Sustainability*, vol. 14, no. 2, 2022.
- [39] J. Y. Park, E. Mistur, D. Kim, Y. Mo, and R. Hoefer, “Toward human-centric urban infrastructure: Text mining for social media data to identify the public perception of COVID-19 policy in transportation hubs.,” *Sustain. cities Soc.*, vol. 76, p. 103524, Jan. 2022.
- [40] X. Yu, Y. Liu, X. Huang, and A. An, “Mining online reviews for predicting sales performance: A case study in the movie domain,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 720–734, 2010.
- [41] H. J. Aleqabie, M. S. Safoq, I. R. Shareef, R. Alsabah, and E. H. Abd, “Sentiment analysis for movie reviews using embedding words with semantic orientation,” *AIP Conf. Proc.*, vol. 2290, no. December, 2020.
- [42] F. Janin Jalal and A. Al-Haj Hussein, “Impact of Customer Relationship Management on Customer Satisfaction in Private Sector Organization,” pp. 1–13, 2018.
- [43] B. Galitsky, “Chatbots for CRM and Dialogue Management,” in *Artificial Intelligence for Customer Relationship Management: Solving Customer Problems*, Cham: Springer International Publishing, 2021, pp. 1–61.
- [44] A. Massaro, N. Magaletti, G. Cosoli, V. Giardinelli, and A. Leogrande, “Text Mining Approaches Oriented on Customer Care Efficiency,” *SSRN Electron. J.*, pp. 0–28, 2022.
- [45] P. Nasa-Ngium, W. S. Nuankaew, and P. Nuankaew, “Analyzing and Tracking Student Educational Program Interests on Social Media with Chatbots Platform and Text Analytics,” *Int. J. Interact. Mob. Technol.*, vol. 17, no. 5, pp. 4–21, 2023.
- [46] O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko, “Mine your own business: Market-structure surveillance through text mining,” *Mark. Sci.*, vol. 31, no. 3, pp. 521–543, 2012.
- [47] M. Pejić Bach, Ž. Krstić, S. Seljan, and L. Turulja, “Text mining for big data analysis in financial sector: A literature review,” *Sustainability*, vol. 11, no. 5, p. 1277, 2019.
- [48] P. Gupta and B. Narang, “Role of text mining in business intelligence,” *Gian Jyoti E-Journal*, vol. 1, no. 2, pp. 1–13, 2012.
- [49] E. Systems and E. Systems, “Business Intelligence in Banking : a Literature Analysis from 2002 to 2013 using Text Mining and Latent Dirichlet Allocation,” no. 351, 2019.
- [50] C.-H. Chou, A. P. Sinha, and H. Zhao, “A text mining approach to Internet abuse detection,” *Inf. Syst. E-bus. Manag.*, vol. 6, no. 4, pp. 419–439, 2008.
- [51] Y. Zhou, X. Wang, and K. F. Yuen, “Sustainability disclosure for container shipping: A text-mining approach,” *Transp. Policy*, vol. 110, pp. 465–477, 2021.
- [52] C.-H. Chung and L.-J. Chen, “Text mining for human resources competencies: Taiwan example,” *Eur. J. Train. Dev.*, vol. 45, no. 6/7, pp. 588–602, Jan. 2021.
- [53] M. Alizadeh, A. Kashef, Y. Wang, J. Wang, G. E. Okudan Kremer, and J. Ma, “Circular economy conceptualization using text mining analysis,” *Sustain. Prod. Consum.*, vol. 35, pp. 643–654, 2023.

- [54] B. Tundys and T. Wiśniewski, “Renewable Energy Supply Chains—Text Mining and Co-Occurrence Analysis in the Context of the Sustainability,” *Energies*, vol. 16, no. 12, 2023.
- [55] M. Işık and H. Dağ, “The impact of text preprocessing on the prediction of review ratings,” *Turkish J. Electr. Eng. Comput. Sci.*, vol. 28, no. 3, pp. 1405–1421, 2020.
- [56] V. Gurusamy and Kannan S, “Preprocessing Techniques for Text Mining,” *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2014.
- [57] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, “Survey on Text Classification Algorithms: From Text to Predictions,” *Inf.*, vol. 13, no. 2, pp. 1–39, 2022.
- [58] P. Y. Shotorbani, “Text Mining Techniques for Analyzing Unstructured,” Texas State University, 2016.
- [59] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, “Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations,” *Organ. Res. Methods*, vol. 25, no. 1, pp. 114–146, 2022.
- [60] M. Saad, “The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification,” The Islamic University-Gaza, 2010.
- [61] S. Pradha, M. N. Halgamuge, and N. Tran Quoc Vinh, “Effective text data preprocessing technique for sentiment analysis in social media data,” *Proc. 2019 11th Int. Conf. Knowl. Syst. Eng. KSE 2019*, pp. 1–8, 2019.
- [62] H. K. Kim, H. Kim, and S. Cho, “Bag-of-concepts: Comprehending document representation through clustering words in distributed representation,” *Neurocomputing*, vol. 266, pp. 336–352, 2017.
- [63] Y. Zhang, M. Chen, and L. Liu, “A review on text mining,” in *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2015, pp. 681–685.
- [64] B. Parlak, “A novel feature and class-based globalization technique for text classification,” *Multimed. Tools Appl.*, pp. 1–26, 2023.
- [65] O. Bridal, “Named-entity recognition with BERT for anonymization of medical records.” 2021.
- [66] Y. B. Kim, B. Snyder, and R. Sarikaya, “Part-of-speech taggers for low-resource languages using CCA features,” *Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process.*, no. September, pp. 1292–1302, 2015.
- [67] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv Prepr. arXiv1810.04805*, 2018.
- [68] G. A. Al-Sultany and H. J. Aleqabie, “Enriching Tweets for Topic Modeling via Linking to the Wikipedia,” *Int. J. Eng. Technol.*, vol. 7, no. 19, pp. 144–150, 2018.
- [69] G. A. Al-Sultany and H. J. Aleqabie, “Events Tagging in Twitter Using Twitter Latent Dirichlet Allocation,” *Int. J. Eng. Technol.*, vol. 7, no. 4.19, pp. 884–888, 2018.
- [70] A. Bello, S. C. Ng, and M. F. Leung, “A BERT Framework to Sentiment Analysis of Tweets,” *Sensors*, vol. 23, no. 1, 2023.
- [71] S. N. inglin Shao and Beijing, “Fine-Grained Sentiment Analysis Based on Hierarchical Attention Networks,” *Comput. Sci. Appl.*, vol. 09, no. 11, pp. 2143–2153, 2019.
- [72] O. M. Al-Janabi, M. K. Ibrahim, A. Kanaan-Jebna, O. M. Alyasiri, and H. J. Aleqabie, “An improved Bi-LSTM performance using Dt-WE for implicit aspect extraction,” in *2022 International Conference on Data Science and Intelligent Computing, ICDSIC 2022*, 2022, pp. 14 – 19.
- [73] A. T. Al-Tuama and D. A. Nasrawi, “Intent Classification Using Machine Learning Algorithms and Augmented Data,” *2022 Int. Conf. Data Sci. Intell. Comput. ICDSIC 2022*, no. Icdsic, pp. 234–239, 2022.
- [74] J. Zhang *et al.*, “Graph convolutional network-strengthened topic modeling for scientific papers,” *Proc. - 2021 IEEE Int. Conf. Smart Data Serv. SMDs 2021*, pp. 24–32, 2021.
- [75] R. Egger and J. Yu, “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts,” *Front. Sociol.*, vol. 7, no. May, pp. 1–16, 2022.
- [76] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.
- [77] P. Laban, L. Dai, L. Bandarkar, and M. A. Hearst, “Can Transformer Models Measure Coherence in Text? Re-Thinking the Shuffle Test,” *ACL-IJCNLP 2021 - 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, vol. 2, pp. 1058–1064, 2021.
- [78] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “{BERT}: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
- [79] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual BERT?,” *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 4996–5001, 2020.
- [80] T. Bolukbasi, K. W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings,” *Adv. Neural Inf. Process. Syst.*, pp. 4356–4364, 2016.

- [81] M. Lamba and M. Madhusudhan, *Text Mining for Information Professionals*. 2022.
- [82] A. T. Al-Tuama and D. A. Nasrawi, "A Survey on the Impact of Chatbots on Marketing Activities," *2022 13th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2022*, no. October, 2022.
- [83] C. Zhang et al., "A complete survey on generative ai (aigc): Is chatgpt from gpt-4 to gpt-5 all you need?," *arXiv Prepr. arXiv2303.11717*, 2023.
- [84] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [85] X. Liu et al., "Large Language Models are Few-Shot Health Learners," *arXiv Prepr. arXiv2305.15525*, 2023.
- [86] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. M. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 02, pp. 52–58, 2021.
- [87] W. Zhai, X. Bai, Y. Shi, Y. Han, Z.-R. Peng, and C. Gu, "Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs," *Comput. Environ. Urban Syst.*, vol. 74, pp. 1–12, 2019.
- [88] D. J. Mohammed and H. J. Aleqabie, "The Enrichment Of MVSA Twitter Data Via Caption-Generated Label Using Sentiment Analysis," in *2022 Iraqi International Conference on Communication and Information Technologies (IICCIT)*, 2022, pp. 322–327.
- [89] J. K. Buser, Y.-J. Cheng, and R. A. McLaughlin, "Thematic Analysis," *Reimagining Res. Engag. Data, Res. Progr. Eval. Soc. Justice Couns.*, p. 153, 2023.
- [90] D. J. Mohammed and H. J. Aleqabie, "Context-Based Visual Sentiment Analysis for Social Media Data," in *2022 International Conference on Artificial Intelligence of Things (ICAIoT)*, 2022, pp. 1–5.
- [91] L. Tan et al., "Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space-air-ground integrated intelligent transportation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2830–2842, 2021.
- [92] B. Müller, L. J. Castro, and D. Rebolz-Schuhmann, "Ontology-based identification and prioritization of candidate drugs for epilepsy from literature," *J. Biomed. Semantics*, vol. 13, no. 1, pp. 1–18, 2022.
- [93] G.-K. J. Li, C. V Trappey, A. J. C. Trappey, and A. A. S. Li, "Ontology-based knowledge representation and semantic topic modeling for intelligent trademark legal precedent research," *World Pat. Inf.*, vol. 68, p. 102098, 2022.
- [94] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge Graphs: Opportunities and Challenges," *Artif. Intell. Rev.*, 2023.
- [95] P. A. Bonatti, S. Decker, A. Polleres, and V. Presutti, "Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web," *Rep. from Dagstuhl Semin.*, vol. 8, no. 09, p. 18371, 2018.
- [96] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, 2022.
- [97] Y. A. A. S. Aldeen, M. Salleh, and M. A. Razzaque, "A comprehensive review on privacy preserving data mining," *Springerplus*, vol. 4, no. 1, pp. 1–36, 2015.
- [98] M. Kayaalp, "Patient privacy in the era of big data," *Balkan Med. J.*, vol. 35, no. 1, pp. 8–17, 2018.
- [99] N. Jindal and B. Liu, "Analyzing and detecting review spam," in *Seventh IEEE international conference on data mining (ICDM 2007)*, 2007, pp. 547–552.
- [100] M. Salehan and D. J. Kim, "Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics," *Decis. Support Syst.*, vol. 81, pp. 30–40, 2016.