ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 11 (2024) No. 6, pp. 198-218
DOI:10.33168/JLISS.2024.0612

Text Anomaly Detection Advancements, Challenges and Pathways: A Systematic Literature Review

Shakirat Oluwatosin Haroon-Sulyman[1], Siti Sakira Kamaruddin[2], Farzana Kabir Ahmad[2]

[1]Department of Information and Communication Science, Faculty of Communication and Information Sciences, University of Ilorin, Ilorin, Nigeria

[2]School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia

sulyman.sh@unilorin.edu.my.ng (Corresponding Author); sakira@uum.edu.my; farzana58@uum.edu.my

Abstract. This systematic literature review synthesized 64 studies from 2017–2023 to analyze recent advances in anomalous text detection methods and applications. The analysis revealed increasing adoption of deep learning methods, especially transformer architectures, given their contextual modeling capabilities. However, fundamental data scarcity and evaluation challenges persist, necessitating tailored datasets, optimized feature engineering, and contextual performance metrics aligned to domain objectives. While increasingly utilized across documents, social media, financial transactions and health sectors, tailoring detection models to capture niche language complexities remains vital. Maintaining rigorous benchmarking, enhancing trainability on scarce anomalies using generative adversarial approaches and emphasizing model transparency emerge as priorities to balance text analytics innovation with ethical application.

Keywords: Text data, Text Anomaly Detection (TAD), text outlier, Natural Language Processing (NLP), supervised learning, unsupervised learning, Transformer Language Model (TLM), generative approach

1. Introduction

Natural language processing (NLP) has become an essential aspect of computer science and artificial intelligence, showing tremendous growth over recent years (Chowdhary, 2020). NLP plays a vital role in information extraction, text classification, sentiment analysis, and information retrieval (Gong et al., 2020; Kumar & Bhatia, 2020). One crucial but less often discussed application of NLP is text anomaly detection (TAD), which focuses on finding unusual patterns or rare occurrences within text data (Jafari, 2022).
Research interest in TAD has grown, indicating its increasing importance across different sectors and purposes. TAD can provide beneficial information for social media spam detection, fishing fraudulent financial transactions, identifying mistakes in healthcare records, and even mapping regulatory tools based on text data uses. This approach helps in drawing valuable insights from large text datasets, supporting better decision-making in many fields across government departments, businesses, organizations and corporations (Fisichella, 2021; Jayadurga et al., 2023; Santiso et al., 2020; Schulte et al., 2022; Suyanto et al., 2022; Wallace & Kecahdi, 2019).

TAD is a somewhat dynamic and complex endeavour due to the challenges in developing an effective model posed by the evolving nature of language, including introducing new words, phrases, slang, terminologies, and languages (Schmitt & Spinosa, 2018). Researchers have explored approaches ranging from machine learning to complex deep learning algorithms to address this challenge. These approaches aim to effectively identify and differentiate between normal and anomalous instances within a dataset (Fisichella, 2021; Rose et al., 2020). Nevertheless, these approaches often struggle due to the need for detailed annotated datasets, the rarity of anomalies in the data, and issues with the datasets not fully representing the diversity of real-world text. These obstacles make it challenging to develop and apply effective anomaly detection models (Bobur, 2020; Cichosz, 2020; Seki et al., 2022; Serra et al., 2018).

The primary objective of identifying anomalies in textual data is to highlight patterns that deviate significantly from normal behaviour. The identified anomalies can help resolve issues before they escalate, assisting decision-makers in making firm and informed decisions based on a clearer understanding of the data (Fisichella, 2021; Rose et al., 2020). As seen generally in research, the typical approach for a TAD task can be grouped into dataset collection, pre-processing of collected data, feature extraction, and anomaly detection method. The anomaly detection methods fall into machine and deep learning categories, which can be further divided into three broad methods: supervised, unsupervised, and semi-supervised. Bobur (2020) demonstrated the supervised method for identifying anomalies in judicial practice, which solely depends on classifying data by relying on labels. This method's advantage is learning and generalising patterns to make predictions. Its major challenge is that it relies solely on labels for data training. Going forward, the unsupervised methods offer a unique ability to understand the underlying semantics of texts without relying on labels (Seki et al., 2022). However, its disadvantages lie in the challenge of scarce anomalous texts in the datasets available for training, and this poses a distinctive hindrance in effectively modeling and detecting anomalies (Cichosz, 2020). And semi-supervised is a combination of both supervised and unsupervised, which uses a small amount of labeled and large unlabeled data for data training. The disadvantage of the semi-supervised method is its inaccuracy when the labeled data does not represent the entire dataset (Serra et al., 2018). It is observed that the selected studies mostly use supervised methods, which are constrained by their reliance on labeled data.

However, dependence on labels for natural language texts is time-consuming and challenging to obtain due to the high dimensionality and complex nature of texts (Balouchzahi et al., 2023; Bhattarai et al., 2022; Wallace & Kecahdi, 2019). The majority of texts are free-form and lack predefined categories or structured patterns (Bhattarai et al., 2021) which makes the use of conventional methods impractical and difficult to extract relevant data features required for anomalous detection, as such, the need for robust and dynamic

text anomaly detection methods becomes more pronounced. Some studies have investigated unsupervised methods that offer a unique ability to understand the underlying semantics of texts without the need for explicit annotations. For instance, (Schulte et al., 2022) presented a deep learning-based method for identifying outliers in short text data. The model clusters data quickly and efficiently. In another study, (Mu et al., 2021) presented a deep learning-based model to handle high-dimensional sparse texts from real-world textual data. The deep learning-based model called TADNET was based on unlabeled data. The results showed significant improvement in detecting anomalous unstructured text data and concluded that deep networks can efficiently capture semantic information from textual data. However, studies on text anomaly detection using unsupervised methods remain relatively underexplored. Given these challenges, this paper aims to systematically review the latest developments in TAD methods, focusing mainly on deep learning techniques like transformers and Generative Adversarial Networks (GANs). We examine how these methods tackle limited data issues and real-world applications' complexity. This review is motivated by the ongoing advancements in TAD, highlighting the need to examine current methods, techniques, applications, challenges, and directions for future research. Thus, this review is organized around four key research questions:

Research Question 1: What are the basic concepts associated with detecting anomalies in text data?

Research Question 2: What are the current methods for identifying features in anomalous text data?

Research Question 3: What anomaly detection methods and application areas are common for detecting anomalies in text data?

Research Question 4: What datasets and metrics are used to evaluate TAD models?

The findings presented in this SLR will provide researchers and readers with a detailed understanding of the current state of TAD research, including basic concepts, challenges, and opportunities for future research work on textual data anomaly detection. The remaining part of this paper is structured into Section 2, which presents the literature review. Section 3 is on research methodology. Section 4 is data synthesis and analysis. Section 5 presents the results. Section 6 discusses the findings, future recommendations, and suggestions for further work. Finally, Section 7 presents the conclusion.

2. Literature Review

2.1 Text Anomaly Detection: Concepts and Applications

Text anomaly detection, interchangeably known as novelty or outlier detection, identifies instances within textual data that deviate significantly from the norm. This identification process is essential for recognizing fraud, ensuring cybersecurity, analyzing social media, and managing healthcare data effectively. A core component of TAD is extracting features from text, such as word frequencies, syntactic patterns, and semantic relationships, which are instrumental in distinguishing anomalies (Ishara Amali & Jayalal, 2020; Yu et al., 2021). The evolution of machine learning algorithms, particularly deep learning, has markedly enhanced the capabilities of anomaly detection systems.

Initial methods predominantly utilized statistical techniques, that assumed a specific data distribution (Lazhar, 2018; Zhu & Zhang, 2017). However, these methods struggled with the dynamic nature of large datasets. This led to a shift towards more adaptable machine learning methods. Support Vector Machines (SVM) and, more recently, methods, deep learning methods like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) offered new pathways by employing supervised learning with feature extraction techniques like Bag-of-Words (BoW), word2vector (word2vec), document2vector (doc2vec), Term Frequency- Inverse Document Frequency (TF-IDF) and N-grams. However, many of these feature extraction techniques are handcrafted word embedding, requiring a manual definition of features (Pizurica & Tomovic, 2020). Additionally, these methods typically rely heavily on labeled data for training,

which can be time-consuming and limit their ability to capture complex relationships within text data (Bhattarai et al., 2021).

The advent of unsupervised deep learning offered promising alternatives, capable of discovering complex patterns and relationships within text without the need for labels. Schulte et al. (2022) and Amplayo et al. (2018) highlighted the effectiveness of clustering-based approaches in grouping texts by their distinct features to detect anomalies. Despite the advantages, these unsupervised methods sometimes fail to capture nuanced semantic connections and explain their findings, particularly in scenarios with sparse anomaly data samples.

To overcome these hurdles, the research community has been proactive. Mu et al. (2021) demonstrated Generative Adversarial Networks (GANs), which can enhance training data quality by generating realistic samples to improve training data diversity. Further, the integration of transformer-based language models, as described by Kumari et al. (2021) and Ghosal et al. (2022), marks a significant leap forward. These models can capture long-range dependencies in text, which is particularly beneficial for addressing the linguistic variability encountered across different domains, leading to better context understanding for anomaly detection.

2.2 Limited Data Availability

Limited data availability remains a significant hurdle for text anomaly detection, hindering the training of robust machine learning and deep learning models. Insufficient data can lead to overfitting, where the model learns specific patterns in the training data but fails to generalize to unseen real-world data. Limited data may not adequately represent the diverse range of patterns present in real-world texts, compromising the accuracy of anomaly detection,

Balouchzahi et al. (2023) emphasized the critical role of large training data, particularly labeled data, for supervised learning methods in text anomaly detection. Supervised learning has the potential for high accuracy, but depends heavily on a sufficient amount of labeled data (Ghosal, et al. 2018). However, in real-world scenarios with limited labeled data, supervised methods face challenges such as scarce labeled anomalous data, expensive labeling processes, difficulty adapting to new data, and time consuming (Al-Adhaileh et al., 2022; Bhattarai et al., 2021).

These limitations particularly on scarcity of labeled anomaly data and inadequate representation of complex patterns, highlight the potential of unsupervised learning methods. Methods like Generative Adversarial Networks (GANs) and Transformers offer alternatives for text anomaly detection in scenarios with limited real-world data availability.

GANs can address the issue of limited data by generating synthetic anomaly data (Mu et al., 2021). This synthetic data can supplement the limited real-world data and potentially improve the training process without extensive manual labeling. This ability to generate synthetic data is particularly valuable for text anomaly detection tasks where labeled anomaly data is scarce. On the other hand, Transformers language models are pre-trained on massive amounts of diverse text data, it can capture intricate semantic relationships within text (Seki et al., 2022). This ability allows them to identify deviations from normal text patterns, even with limited training data, improving anomaly detection performance.

2.3 Evaluation Metrics

Evaluation metrics assess the performance of a text anomaly detection model on its effectiveness. Effectively evaluating the performance of a text anomaly detection model is crucial for understanding its strengths and weaknesses. It is important to note that different methods are used to evaluate anomaly detection models, and the choice of evaluation metrics depends on the specific characteristics of the text data and the nature of anomalies being detected in a particular domain. The primary goal is to compare the

performance of different anomaly detection models (Sepúlveda-Torres et al., 2021). Supervised methods have some common metrics, including precision, recall, F1-score, accuracy, and area under the receiver operating characteristic curve, which are ideally suited for training data labeled normal and anomalous instances. However, evaluation becomes more challenging in unsupervised anomaly detection methods where labeled data is unavailable (Ruff et al., 2019). Identifying true anomalies becomes more difficult, some labeled examples might still be required for evaluation, despite not being used for training the model (Mu et al., 2021). Schulte et al. (2022) used reconstruction error as an unsupervised evaluation metric for financial record outlier detection. Some others used a combination of Mean Square Error (MSEs) with correlation coefficient to gain insight to the quality of anomaly detection in texts and how they align with human intuition (Seki et al., 2022).

While unsupervised evaluation metrics offer valuable tools for identifying potential anomalies, especially in scenarios with limited labeled data, choosing the most appropriate ones requires careful consideration of various factors. These factors include the specific application domain, the nature of the targeted anomalies, and the limitations of each metric itself.

3. Research Methodology

3.1. Research Design

This systematic literature review adopted a research design following the guidelines outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) checklists and flow diagram (Page et al., 2021). Figure 1 depicts the overall phases of the search strategy. The initial phase established the research protocol, defined the research questions, specified the search keywords, and determined the appropriate search databases.

Subsequent phases entailed applying predetermined inclusion and exclusion criteria to screen and select relevant literature systematically. This approach promotes transparency, consistency, accountability, and integrity in systematic reviews. It further emphasizes careful planning and documentation of the review process. The final stage synthesized and analyzed the results obtained from the selected articles, culminating in a comprehensive review adhering to established guidelines for systematic reviews and meta-analyses.

3.2. Database Selection

The bedrock of a high-quality review lies in the comprehensiveness and rigour of the data sources utilized for identifying relevant articles. This requires an exhaustive search across various academic databases to ensure a thorough examination of existing literature. For this review, reputable bibliography databases which include Scopus, IEEE Xplore, and Science Direct were chosen. These databases are widely recognized journal articles or conference papers that are highly indexed and reliable, needed to capture a comprehensive and high-quality collection of literature pertinent to the research topic.

3.3. Search Keyword

A carefully crafted combination of keywords and search queries was utilized to perform a comprehensive literature search for this systematic review, leveraging Boolean operators (AND, OR) to combine terms related to the target domain. The search string employed was as follows: "text" AND ("anomaly" OR "outlier" OR "deviation" OR "novelty") AND ("detection" OR "identification"). This strategically constructed query ensured that the retrieved studies were directly relevant to textual anomaly detection, encompassing various related terms and concepts.

3.4. Inclusion Criteria

To guarantee a fair selection process and avoid favoring specific articles, a clear criterion for paper inclusion

was established to guide the selection process objectively and eliminate potential biases. These inclusion criteria are detailed in Table 1.

Hence, the specified search queries across the various databases yielded an initial pool of 650 relevant articles in English, covering the subjects of interest. The search encompassed all available online publications in the domain of this review from 2017 to 2023. The subsequent section outlines the process to methodically screen this pool of 650 articles methodically, ensuring that only those meeting the established criteria and aligning with the study's objectives were included in the final analysis.

Table 1: Inclusion Criteria

S/N	Inclusion Criteria
1.	The study was published between 2017-2023
2.	The study is written in English language
3.	The study focuses on Textual data domain and specifically on Anomaly Detection
4.	The study is either a journal article, conference paper, or book chapter

3.5. Exclusion Criteria

The initial database keyword search yielded 650 articles. To identify the most relevant studies for the review, a systematic elimination procedure was employed, consisting of the following stages:

i. The first stage involved excluding articles based on unrelated titles, and duplicates from various sources. This initial filtering reduced the pool from 650 to 547 articles.

ii. In the second stage, the titles, keywords, and abstracts of the remaining 547 articles were examined, further narrowing down the selection to 169 potentially relevant papers.

iii. The third stage entailed a more focused assessment, eliminating articles not specifically addressing text data anomaly, novelty, outlier, or deviation detection techniques. This step reduced the count from 169 to 80 articles.

iv. Finally, the full texts were thoroughly read, and articles that deviated from the study's core objectives were excluded. This final screening process resulted in a refined set of 64 highly pertinent articles.

The systematic elimination procedure is visually represented in Figure 1, which depicts the sequential flow of the screening process employed to identify the most relevant literature for the study.

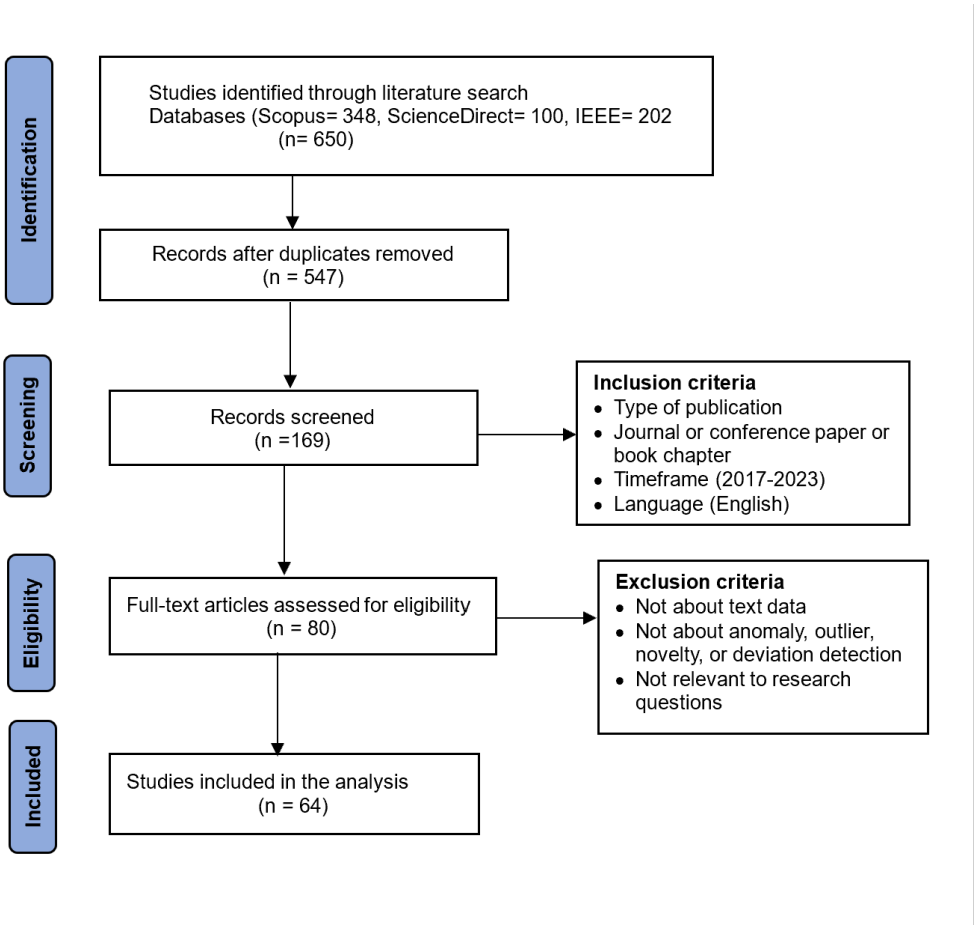

Fig. 1: PRISMA search methodology

Table 2 provides a summary of the databases explored and the respective number of articles until the final 64 obtained from each source. It shows the percentage of the articles sourced from each academic database. It reveals that Scopus has the highest number of related articles with 35 (55%) papers, followed by IEEE with 23 (36%), and the least paper contribution is by Science Direct with 6 papers, which represents 9% of the entire papers reviewed for this systematic literature review. The next section summarizes and reports the extracted data.

Table 2: Search databases and number of articles

S/N	Article Sources	URL	No. of Articles	Percentage (%)
1.	Scopus	https://www.scopus.com/	35	55
2.	IEEE	http://www.ieeexplore.ieee.org/	23	36
3.	Science Direct	http://www.sciencedirect.com/	6	9
		Total =	64	100

4. Data Synthesis and Analysis

This section aims to synthesize and summarize information from the included studies by answering the research questions presented earlier. It aims to enable readers and researchers to identify the current state of the art on textual anomalous detection.

A. Research Question 1: What are the Concepts Related to Textual Data Anomaly Detection?

Several concepts are related to textual anomaly detection, including Natural Language Processing (NLP) techniques, datasets (collection and preprocessing), anomalies in the text domain (feature extraction and anomaly detection methods), application domains, and evaluation metrics.

B. Research Question 2: What are the Current Feature Extraction Techniques for Anomalous Textual Data Detection?

Based on the literature reviewed, the widely used FE technique is word embedding, which captures the semantic relationship between words (Dynich & Wang, 2017; Jayadurga et al., 2023; Mohotti & Nayak, 2020; Yu et al., 2021). Character-level focus on obtaining rich morphological information within words and segmenting words into sub-words or characters. Contextual embedding aims to capture word meanings by considering the context of words in a document and offers models that can be fine-tuned for specific tasks (Ghosal et al., 2022; Seki et al., 2022). Hybrid techniques (Kumari et al., 2021; Sepúlveda-Torres et al., 2021) based on combining more than one technique have also been used to enhance TAD robustness based on data characteristics. Additionally, some studies explored alternative techniques like spherical and graphical-based embedding. Based on the literature observation, 51 out of the 64 reviewed studies used either word, character-level, or contextual embedding techniques, with 47% on word embedding, 3% for character-level, 6% for contextual level, 22% for hybrid and 21% others used alternative techniques. Hence, word embedding appears to be the most used technique in the studied years (2017-2023). Figure 1 shows the distribution of the various categories applied in the studies.

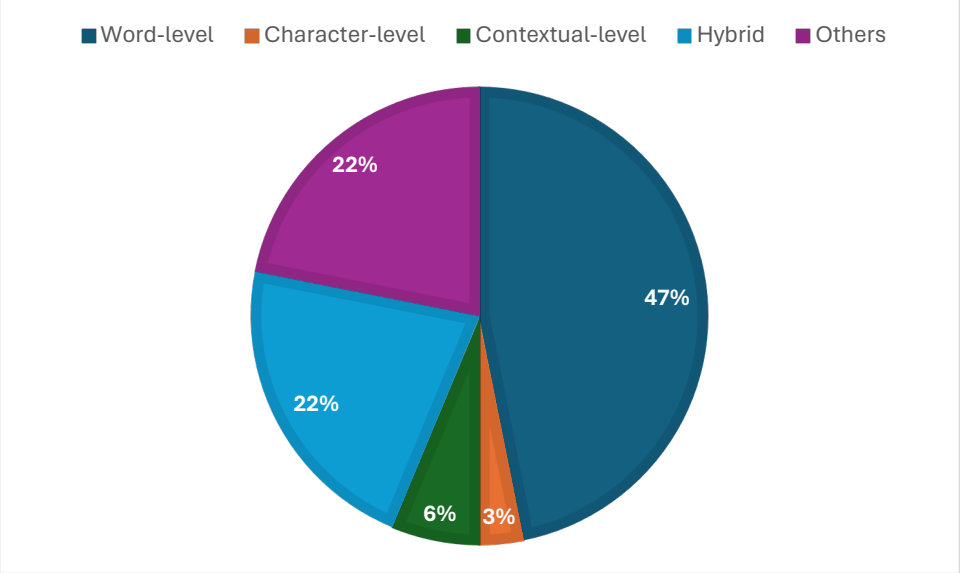

Fig. 3: Distribution of FE technique categories used in selected studies.

- Word embedding techniques

Word embedding is a widely chosen FE technique in the TAD domain, crucial across various NLP tasks. It maps words as a single vector representation (Fisichella, 2021). An example of its application area is identifying anomalous flight report narratives (Rose et al., 2020). Based on (Bendella et al., 2018; Wang et al., 2017) word embedding techniques have been used in TAD research for more than five years. The goal of this technique is to capture semantic meaning with the incorporation of frequency-based information; particularly typical of unstructured texts with a lack of predefined structures (Al-Adhaileh et al., 2022; Doboli et al., 2020; Mohotti & Nayak, 2020). It has evolved with the use of several methods like a bag of words (BoW), word2vec, Glove, and Term Frequency-Inverse Document Frequency (TF-IDF) (Jayadurga et al., 2023). Challenges with this category are the inability to capture word morphology and OOV (Out of Vocabulary) extension, where they cannot handle novel words outside the training data.

- Character-level techniques

To handle morphologically rich languages and OOV words, character levels are used to offer a nuanced approach to linguistic representations. It represents words or characters as vectors in a continuous space, that even when encountered with novel words not present in the training data they can still provide meaningful representations based on the characters they contain (Mohaghegh & Abdurakhmanov, 2021). An example of its application is identifying outliers in pseudepigraphic texts (Koppel & Seidman, 2018). Character level appears the least used in the selected studies with methods such as n-gram (Ghosal et al., 2018). The trend shows the combination of character-level and word embedding using methods like unigram and TF-IDF to handle words that do not appear in the vocabulary (OOV) and capture rich textual patterns respectively for an improved performance (Balouchzahi et al., 2023; Santiso et al., 2020). However, dealing with unseen characters, especially in diverse language sets is a challenge to them.

- Contextual techniques

Unlike previously seen techniques that map words to the same vector, contextual techniques are required to map words to different embedding vectors depending on their context (Ghosal et al., 2022). Contextual embedding aims to manage the varying word meanings based on their contexts. The examined studies showed its various applications with remarkable results in detecting anomalous texts in areas like business sentiment, and judiciary (Bobur, 2020; Seki et al., 2022). As evident in the selected studies, there has been an observable growth in its use with already trained Transformer Language Models (TLMs) such as Bidirectional Encoder Representations from Transformers (BERT) (Ghosal et al., 2022; Kumari et al., 2021; Manolache et al., 2021). While it focuses on capturing contextual nuances, it faces challenges acquiring labeled data for fine-tuning domain-specific data.

Table 3: Type of FE techniques

List of studies	FE technique
(Asiri et al., 2022; Ghosal et al., 2021; Mu et al., 2021)	GloVe
(Lazhar, 2018; Thomas et al., 2019; Wang et al., 2020)	doc2vec
(Bendella et al., 2018; Bhattarai et al., 2021; Fisichella, 2021; Mei et al., 2018)	BoW
(Al-Adhaileh et al., 2022; Amorim et al., 2019; Efrizoni et al., 2022; Serra et al., 2018; Wallace & Kecahdi, 2019; Wang et al., 2017)	word2vec
(Koppel & Seidman, 2018)	ngram
(Bose et al., 2019; Doboli et al., 2020; Dynich & Wang, 2017; Ghosal et al., 2018; Kannan et al., 2017; Kokatnoor & Krishnan, 2020; Mohotti & Nayak, 2020; Wang et al., 2018; Wang & Chen, 2017; Zhu & Zhang, 2017)	TF-IDF
(Bobur, 2020; Ghosal et al., 2022; Seki et al., 2022)	BERT
(Pizurica & Tomovic, 2020; Pv & Bhanu, 2020; Rose et al., 2020)	BoW & TF-IDF
(Cichosz, 2020)	GloVe & BoW
(Nguyen et al., 2019; Schmitt & Spinosa, 2018)	GloVe & word2vec
(Ruff et al., 2019)	GloVe & FastText
(Kumari et al., 2021)	GloVe & BERT
(Walkowiak et al., 2018)	TF-IDF & FastText
(Sepúlveda-Torres et al., 2021)	TF-IDF & RoBERTa
(Yu et al., 2021)	TF-IDF & BERT
(Jayadurga et al., 2023)	TF-IDF & word2vec
(Ghosal, Salam, et al., 2018)	word2vec & ngram
(Balouchzahi et al., 2023)	unigram & TF-IDF

C. Research Question 3: What are the Anomaly Detection Methods and Application Areas for Detecting Anomalous Textual Data?

The quality of the extracted features plays a vital role in the success of an anomaly detection task, thereby making selected methods for FE and AD crucial in textual data analysis. This study identified categories of feature extraction techniques from textual data, including word, character, contextual, hybrid, and other techniques. The nuances captured during feature extraction, such as semantic relationships, syntactic structures, and contextual information become the basis for anomaly detection methods, making it the foundation for anomaly detection tasks. Further, various methods to identify anomalous patterns within textual data were examined from the studies gathered. Anomaly detection methods, ranging from supervised, unsupervised, and semi-supervised, leverage the extracted features to detect deviations from the expected patterns. Supervised methods rely on labels or annotations to train a model, such that the model learns to distinguish between normal and anomalous classes based on predefined labels. Unsupervised methods use unlabeled data, to identify patterns or relationships without labels for data training. And, semi-supervised is a mixture of both supervised and unsupervised methods. It is trained with a small amount of labeled data and many unlabeled data. Based on the reviewed studies, one or more methods have been applied in various application areas. Table 2 summarizes the types of methods used in various categories and areas applied with Figure 2 showing the frequency of various application domains. Despite few of the studies considered for this review were not included in Table 2 because neither of the methods or application areas was explicitly stated, a thorough observation of the table revealed that the majority of the selected studies fall under the supervised category with various machine and deep learning methods used either as a single method or combination of more than one (hybrid). The challenge with the supervised method is their reliance on labeled data for training which is challenging because anomalies are infrequent and evolve.

The use of unsupervised methods which doesn't rely on labels is revealed in Table 2 with the majority using machine learning methods such as one-class SVM (Seki et al., 2022) and K-means (Rose et al., 2020), and a few deep learning methods like autoencoder (Amplayo et al., 2018) and GAN (Mu et al., 2021). It disclosed that the traditional unsupervised machine learning methods struggle to handle relationships and nuanced patterns in natural language. The semi-supervised, rarely used in the selected studies is challenged with scarcely labeled anomalies that hinder the model's ability to generalize well to real-life scenarios.

Table 4: Categories, types, and application areas of anomaly detection methods used.

Categories	Method type	Application area	References
Supervised	MLP	Document	(Ghosal et al., 2022; Doboli et al., 2020)
	CNN	Document	(Ghosal et al., 2018)
	CNN +BiLSTM	Social media	(Al-Adhaileh et al., 2022)
	BiLSTM	Social media	(Kumari et al., 2021)
	BiLSTM	Document	(Ghosal et al., 2021)
	CNN	SM	(Muppudathi & Krishnasamy, 2023)
	Convolutional BiLSTM	SM	(Yu et al., 2021)
	RNN	SM	(Huang et al., 2018)
	KNN	Document	(Fouche et al., 2020)
	LR +XGB	Document	(Bobur, 2020)
	KNN	Document	(Mohotti & Nayak, 2020)
	KNN+SVM+NB	Social media	(Ishara Amali & Jayalal, 2020)
	KNN+RF	Social Network	(Pv & Bhanu, 2020)
	LR+NB	Document	(Pizurica & Tomovic, 2020)
	KNN	Social Network	(Hassan et al., 2020)
	LSTM+CNN	Financial	(Thomas et al., 2019)
	LSTM	Health	(Wallace & Kecahdi, 2019)
	LSTM+CNN	SN	(Nguyen et al., 2019)
	RF	Document	(Ghosal, Salam, et al., 2018)
	SVM+DT+MLP	Document	(Walkowiak et al., 2018)
	SVM	Document	(Koppel & Seidman, 2018)
	LR+RF+SVM	SM	(Balouchzahi et al., 2023; Jayadurga, 2023)
Unsupervised	Autoencoder	Financial	(Schulte et al., 2022)
	One-class SVM	Business	(Seki et al., 2022)
	GMM	Health	(Fisichella, 2021)
	K-means, GAN	Social media	(Mu et al., 2021)
	K-means	Document	(Rose et al., 2020)
	Fuzzy c-means	Social Network	(Sarna & Bhatia, 2020)
	One-class SVM	Document	(Ruff et al., 2019; Wang et al., 2017)
	LOF	Document	(Shi et al., 2019; Wang et al., 2018)
	DBSCAN	SN	(A. Bose et al., 2019)
	Fuzzy k-means	Document	(Lazhar, 2018)
	One-class SVM+Kmeans	SM	(Cichosz, 2020)
	IF+LOF+Autoencoder	SM	(Amorim et al., 2019)
	Autoencoder+K-means	Document	(Mei et al., 2018)
	Autoencoder	Document	(Amplayo et al., 2018)
Semi-Supervised	CNN+GAN	SN	(Fadhel & Nyarko, 2019)
	LSTM+GMM	Document	(Serra et al., 2018)
	CNN+DBSCAN	Document	(Schmitt & Spinosa, 2018)
	LSTM+VAE	SM	(Shaalan et al., 2021)

Table 5: Summary of Machine Learning Methods Used in the Selected Studies

Method	References
KNN	(Fouche et al., 2020)
LR +XGB	(Bobur, 2020)
KNN	(Mohotti & Nayak, 2020)
KNN+SVM+NB	(Ishara Amali & Jayalal, 2020)
KNN+RF	(Pv & Bhanu, 2020)
LR+NB	(Pizurica & Tomovic, 2020)
KNN	(Hassan et al., 2020)
RF	(Ghosal, Salam, et al., 2018)
SVM	(Koppel & Seidman, 2018)
LR+RF+SVM	(Balouchzahi et al., 2023; Jayadurga, 2023)
One-class SVM	(Seki et al., 2022)
GMM	(Fisichella, 2021)
K-means	(Mu et al., 2021)
Fuzzy c-means	(Rose et al., 2020)
One-class SVM	(Sarna & Bhatia, 2020)
LOF	(Ruff et al., 2019; Wang et al., 2017)
DBSCAN	(Shi et al., 2019; Wang et al., 2018) (Bose et al., 2019)
Fuzzy k-means	(Lazhar, 2018)
One-class SVM+Kmeans	(Cichosz, 2020)

Table 6: Summary of Deep Learning methods used in the selected studies.

Method	Reference
MLP	(Ghosal et al., 2022; Doboli et al., 2020)
CNN	(Muppudathi & Krishnasamy, 2023; Ghosal et al., 2018)
CNN +BiLSTM	(Al-Adhaileh et al., 2022; Yu et al., 2021)
BiLSTM	(Ghosal et al., 2021; Kumari et al., 2021)
RNN	(Huang et al., 2018)
LSTM+CNN	(Thomas et al., 2019; Nguyen et al., 2019)
LSTM	(Wallace & Kecahdi, 2019)
Autoencoder	(Amplayo et al., 2018; Schulte et al., 2022;)
CNN+GAN	(Fadhel & Nyarko, 2019)
LSTM+VAE	(Shaalan et al., 2021)

The majority of the selected studies identified the application domains, this is shown in Figure 3 with 29 studies that focused on the document domain, 23 on social media or social network sites,

5 on business or financial, and 4 on the health domain. Three selected studies (Bhattarai et al., 2021; Manolache et al., 2021; Sudha & Suguna, 2019) weren't explicit on the application domain.

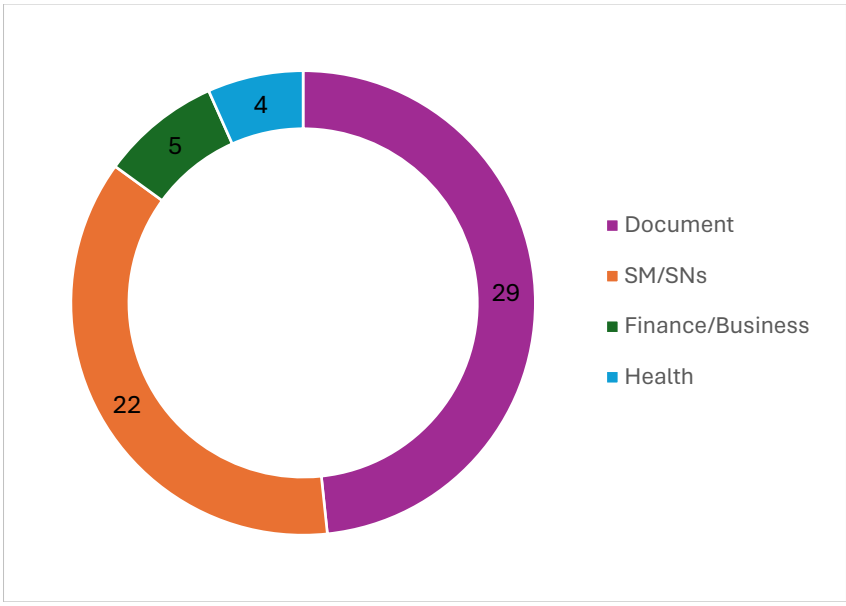

Fig. 4: Count of included publications according to application areas.

D. Research Question 4: What are the Current Datasets, Preprocessing Techniques, and Evaluation Metrics for the Text Anomalous Detection Model?

The nature and type of datasets collected for a TAD task are vital due to the data's unstructured, dynamic, and complex nature. We identified several datasets that apply to real-world scenarios for unstructured text anomaly detection. As shown in Table 6, the summary of the various datasets, preprocessing steps, and evaluation metrics used were provided. The widely used datasets appear to be APSWJ (Associated Press and Wall Street Journal), TAP-DNLD, 20NewsGroup, Reuters, Movie Review (MR), and Yelp which are comprised of a diverse range of topics or news articles in various categories.

Data processing is vital in determining the success or failure of most anomaly detection models. We identified various preprocessing techniques applied to clean the datasets. Preprocessing techniques such as converting words to lowercase, removing stop words, and lemmatization play an important role in refining the collected raw texts to enhance the effectiveness of anomaly detection models. It was observed that preprocessing steps like removing stop words, punctuation marks, and special characters are the widely used preprocessing techniques.

The evaluation metric is the yardstick for measuring the model's performance detecting anomalous texts. We observed that researchers commonly use precision, recall, and AUROC to measure the overall model effectiveness.

5. Results

Having carefully applied the selection procedures as outlined in Section 3, the 64 selected articles were subjected to undergo a systematic review to extract the required information as highlighted in the previous Section. The information is tabulated for further analysis, however, not all the articles contained the desired data items. An analysis of the publication years reveals a growing interest in deep learning methods for text anomaly detection, published between 2019-2023, suggesting a recent surge in research activity in this domain.

5.1. Deep Learning Methods for Text Anomaly Detection

The adoption of deep learning has completely transformed the domain of text anomaly detection, enabling the growth of advanced and efficient approaches to detect anomalies in textual data. According to the reviewed studies, Convolutional Neural Networks (CNNs) have been widely explored across various real-world domains. It has been used to identify document anomalies (Ghosal et al. 2018) and social media (Muppudathi & Krishnasamy, 2023). CNNs have been combined with methods such as Long-short Term Memory (LSTM), and Bidirectional Long-short Term Memory (BiLSTM). For instance, Thomas et al., (2019) developed an LSTM-CNN model for detecting anomalies in financial documents, similarly (Al-Adhaileh et al., 2022) identified strange social media posts. Findings also identified methods like Multilayer Perceptron (MLP) (Doboli et al., 2020; Ghosal et al., 2022), Recurrent Neural Network (RNN) (Huang et al., 2018), Autoencoder (Schulte et al., 2022) which demonstrate their effectiveness in identifying anomalies across various application areas. Findings show that these deep learning methods ensure scalability and automatically learn features from training data, however, they require a large amount of data for training.

Based on the review of the existing studies, it can be deduced that various feature extraction techniques such as word2vec and GloVe coupled with deep learning methods have been deployed to build anomaly detection models. While a significant advantage of deep learning lies in its ability to automatically learn features from data, most deep learning can be susceptible to overfitting, especially with limited training data. However, some algorithms like Autoencoders with unlabeled data can potentially mitigate overfitting. They might struggle with the interpretability of the learned features. Additionally, some methods require large amounts of labeled data for optimal performance.

5.2. Evolving Techniques for Text Anomaly Detection

The application of Transformers for text anomaly detection has gained significant attention in recent years. Studies published from year 2020 (Bobur, 2020; Ghosal et al., 2022; Seki et al., 2022) demonstrate their effectiveness in capturing long-range dependencies and contextual relationships within text data, leading to an improved performance compared to the traditional deep learning methods. Based on the review, transformers prove their ability to understand the context and word meanings more effectively as compared to other techniques like word embeddings. Ghosal et al. (2022) explored the pre-trained Bidirectional Encoder Representations from Transformers (BERT) effectiveness in identifying semantic-level redundancies, leading to better anomaly detection than techniques like GloVe. Similarly, Sepúlveda-Torres et al. (2021) utilized a fine-tuned Robustly optimized BERT (RoBERTa) approach for detecting misinformation in news documents. Kumari et al. (2021) further highlight BERT's ability to handle diverse and dynamic text data. Despite these advantages, a limited use of these models is revealed in the reviewed studies.

5.3. Challenges in Text Anomaly Detection

While deep learning techniques have demonstrated remarkable success in text anomaly detection, their effective implementation and real-world deployment are accompanied by several interrelated challenges and considerations. A fundamental hurdle lies in the limited availability of labeled anomaly data, as anomalies are inherently rare occurrences. This data scarcity can lead to overfitting and hinder the generalizability of models to unseen real-world scenarios. Researchers have explored strategies such as data augmentation, semi-supervised learning, and synthetic data generation using techniques like Generative Adversarial Networks (GANs) to mitigate this issue (Fadhel & Nyarko, 2019; Mu et al., 2021). However, developing specialized datasets tailored to specific anomaly detection tasks within a domain remains crucial. These datasets should accurately reflect the complexities and nuances of the targeted domain's language while ensuring a balanced representation of normal and abnormal data points. Failure to capture these domain-specific characteristics can limit the model's ability to detect subtle deviations or context-specific

anomalies. Effective feature engineering techniques are also essential for transforming raw text into meaningful representations that capture the underlying patterns relevant to anomaly detection. Identifying the most informative features for a specific domain requires domain expertise and extensive experimentation, as the choice of feature engineering techniques can significantly impact model performance and generalizability. Furthermore, model performance evaluation must be carefully considered, as generic metrics like accuracy or F1-score might not align with the desired outcomes and priorities of anomaly detection within a particular domain (Serra et al., 2018). Incorporating domain-specific performance metrics, domain knowledge, and input from subject matter experts becomes crucial to ensure that the developed systems meet real-world deployments' practical requirements and constraints.

6. Discussion

The synthesis of challenges and future directions in textual anomaly detection elucidates critical gaps in literature and practice. The SLR highlights a foundational challenge: the steep learning curve faced by new entrants who must grapple with complex NLP concepts to understand textual anomalies. This task is inherently context-sensitive and nuanced. Despite the SLR's informative nature, there is a pressing need for ongoing engagement with evolving studies to fully assimilate these intricate subjects (Lochter et al., 2018). Feature extraction techniques are critiqued for their inability to capture the full range of linguistic nuances. The prevailing reliance on word embeddings is noted for overlooking finer semantic details essential to anomaly detection. At the same time, character-level and contextual approaches lack in capturing semantic depth and broader contexts, respectively. The uptake of transformer models shows promise, but the field must refine these tools to suit the specific needs of diverse application areas (Ghosal et al., 2022; Kumari et al., 2021; Manolache et al., 2021).

A practical barrier identified is the rarity of anomalies in datasets, which leads to AD methods struggling to generalize to unseen patterns—a problem exacerbated by the limited diversity in current benchmark datasets. This calls for creating richer, more representative datasets that embody the full complexity of real-world textual data (Mohotti & Nayak, 2020; Mu et al., 2021).

The limitations around search completeness and excluded study types suggest that the SLR's scope might not fully reflect the latest and broadest spectrum of research, possibly missing non-indexed studies, non-English literature, or emergent findings from grey literature. This represents a significant limitation as practical innovations might lag or outpace documented research.

Future research directions must contend with these challenges by seeking comprehensive, context-aware, and dynamic methodologies. Innovations in feature extraction must parse linguistic complexity and adapt to specific domain requirements. Datasets must be curated to reflect the real-world diversity and anomaly prevalence, while preprocessing techniques must maintain semantic integrity, avoiding data quality degradation. Evaluation metrics should be chosen for their applicability to specific domains, ensuring that performance evaluations align with real-world needs and not just theoretical models. The critical need for exhaustive searches and including a broader array of study types in SLRs cannot be overstated, to provide a more holistic and up-to-date understanding of the field's cutting edge.

7. Conclusion

This Systematic Literature Review (SLR) meticulously charts the terrain of textual anomaly detection, examining a collection of 64 studies to uncover the strides and sticking points in methods, applications, and datasets within the 2017-2023 timeframe. It underscores the shift towards sophisticated deep learning frameworks, notably transformer-based models, that significantly advance understanding contextual textual nuances, despite the challenges of computational demand and the necessity for diverse training datasets.

The exploration spans various domains, each with tailored requirements, revealing a pressing need for datasets encapsulating real-world complexity and preprocessing that maintains semantic integrity. Furthermore, the SLR critically reflects on the choice of evaluation metrics, advocating for domain-specific metrics over-generalized ones. While comprehensive, the review acknowledges its limitations due to language and database selection biases. It points to the uncharted territories in anomaly detection research—territories rich with potential for future inquiry into more refined, efficient, and contextually aware anomaly detection systems.

Acknowledgement

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UUM/02/1).

References

Al-Adhaileh, M. H., Aldhyani, T. H. H., & Alghamdi, A. D. (2022). Online Troll Reviewer Detection Using Deep Learning Techniques. *Applied Bionics and Biomechanics*, *2022*, 1–10. https://doi.org/10.1155/2022/4637594

Amorim, M., Bortoloti, F. D., Ciarelli, P. M., Salles, E. O. T., & Cavalieri, D. C. (2019). Novelty Detection in Social Media by Fusing Text and Image into a Single Structure. *IEEE Access*, *7*, 132786–132802. https://doi.org/10.1109/ACCESS.2019.2939736

Amplayo, R. K., Hong, S. L., & Song, M. (2018). Network-based approach to detect novelty of scholarly literature. *Information Sciences*, *422*, 542–557. https://doi.org/10.1016/j.ins.2017.09.037

Asiri, Y., Halawani, H. T., Alghamdi, H. M., Hassan, S., Hamza, A., Abdel-khalek, S., & Mansour, R. F. (2022). *applied sciences Enhanced Seagull Optimization with Natural Language Processing Based Hate Speech Detection and Classification*.

Balouchzahi, F., Butt, S., Sidorov, G., & Gelbukh, A. (2023). ReDDIT: Regret detection and domain identification from text. *Expert Systems with Applications*, *225*(December 2022), 120099. https://doi.org/10.1016/j.eswa.2023.120099

Bendella, M., Almuhisen, F., & Quafafou, M. (2018). Geo-Fuzz: Fuzzy-based algorithm for suspicious geo-tagged tweets detection. *IEEE International Conference on Fuzzy Systems*, *2018-July*, 1–7. https://doi.org/10.1109/Fuzz-Ieee.2018.8491533

Bhattarai, B., Granmo, O. C., & Jiao, L. (2021). Measuring the novelty of natural language text using the conjunctive clauses of a tsetlin machine text classifier. *ICAART 2021 - Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, *2*, 410–417. https://doi.org/10.5220/0010382204100417

Bhattarai, B., Granmo, O. C., & Jiao, L. (2022). Word-level human interpretable scoring mechanism for novel text detection using Tsetlin Machines. *Applied Intelligence*, *52*(15), 17465–17489. https://doi.org/10.1007/s10489-022-03281-1

Bobur, M. (2020). Anomaly Detection Between Judicial Text-Based Documents. *IEEE 14th International Conference on Application of Information and Communication Technologies (AICT) | 978-1-7281-7386-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/AICT50176.2020.9368621 Anomaly*.

Bose, R., Dey, R. K., Roy, S., & Sarddar, D. (2019). Analyzing political sentiment using Twitter data. In

Smart Innovation, Systems and Technologies (Vol. 107). Springer Singapore. https://doi.org/10.1007/978-981-13-1747-7_41

Chowdhary, K. R. (2020). Fundamentals of artificial intelligence. In *Fundamentals of Artificial Intelligence*. https://doi.org/10.1007/978-81-322-3972-7

Cichosz, P. (2020). *Unsupervised modeling anomaly detection in discussion forums posts using global vectors for text representation*. 1–28. https://doi.org/10.1017/S1351324920000066

Doboli, S., Kenworthy, J., Paulus, P., Minai, A., & Doboli, A. (2020). A cognitive inspired method for assessing novelty of short-text ideas. *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN48605.2020.9206788

Dynich, A., & Wang, Y. (2017). Analysis of novelty of a scientific text as a basis for assessment of efficiency of scientific activities. *Journal of Organizational Change Management*, *30*(5), 668–682. https://doi.org/10.1108/JOCM-10-2016-0226

Efrizoni, L., Defit, S., & Tajuddin, M. (2022). Hybrid Modeling to Classify and Detect Outliers on Multilabel Dataset based on Content and Context. *International Journal of Advanced Computer Science and Applications*, *13*(12), 550–559. https://doi.org/10.14569/IJACSA.2022.0131267

Fadhel, M. Ben, & Nyarko, K. (2019). GAN Augmented Text Anomaly Detection with Sequences of Deep Statistics. *2019 53rd Annual Conference on Information Sciences and Systems, CISS 2019*. https://doi.org/10.1109/CISS.2019.8693024

Fisichella, M. (2021). Unified approach to retrospective event detection for event- based epidemic intelligence. *International Journal on Digital Libraries*, *22*(4), 339–364. https://doi.org/10.1007/s00799-021-00308-9

Fouche, E., Meng, Y., Guo, F., Zhuang, H., Bohm, K., & Han, J. (2020). Mining text outliers in document directories. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2020-Novem*(Icdm), 152–161. https://doi.org/10.1109/ICDM50108.2020.00024

Ghosal, T., Edithal, V., Ekbal, A., Bhattacharyya, P., Chivukula, S. S. S. K., & Tsatsaronis, G. (2021). Is your document novel? Let attention guide you. An attention-based model for document-level novelty detection. *Natural Language Engineering*, *27*(4), 427–454. https://doi.org/10.1017/S1351324920000194

Ghosal, T., Edithal, V., Ekbal, A., Bhattacharyya, P., Tsatsaronis, G., & Chivukula, S. S. S. K. (2018). Novelty goes deep. A deep neural solution to document level novelty detection. *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, 2802–2813.

Ghosal, T., Saikh, T., Biswas, T., Ekbal, A., & Bhattacharyya, P. (2022). Novelty Detection: A Perspective from Natural Language Processing. *Computational Linguistics*, *48*(1), 77–117. https://doi.org/10.1162/COLI_a_00429

Ghosal, T., Salam, A., Tiwari, S., Ekbal, A., & Bhattacharyya, P. (2018). TAP-DLND 1.0: A corpus for document level novelty detection. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 3541–3547.

Gong, L., Zhang, Z., & Chen, S. (2020). Clinical named entity recognition from Chinese electronic medical records based on deep learning pretraining. *Journal of Healthcare Engineering*, *2020*. https://doi.org/10.1155/2020/8829219

Hassan, N., Abd-Elmegid, L. A., & Helmy, Y. K. (2020). An efficient model for mining outlier opinions.

International Journal of Advanced Computer Science and Applications, *11*(5), 146–153. https://doi.org/10.14569/IJACSA.2020.0110522

Huang, Y., Chung, W., & Tang, X. (2018). A temporal recurrent neural network approach to detecting market anomaly attacks submission type: Short paper. *2018 IEEE International Conference on Intelligence and Security Informatics, ISI 2018*, 160–162. https://doi.org/10.1109/ISI.2018.8587397

Ishara Amali, H. M. A., & Jayalal, S. (2020). Classification of Cyberbullying Sinhala Language Comments on Social Media. *MERCon 2020 - 6th International Multidisciplinary Moratuwa Engineering Research Conference, Proceedings*, 266–271. https://doi.org/10.1109/MERCon50084.2020.9185209

Jafari, A. (2022). *A Deep Learning Anomaly Detection Method in Textual Data*. http://arxiv.org/abs/2211.13900

Jayadurga, R., Veeramakali, T., Sohail, M. A., Balaji, N. A., Kiran Kumar, C., & Sajitha, L. P. (2023). Deep Learning Based Detection and Classification of Anomaly Texts in Social Media. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(4s), 78–89.

Kannan, R., Woo, H., Aggarwal, C. C., & Park, H. (2017a). *Outlier Detection for Text Data : An Extended Version*. http://arxiv.org/abs/1701.01325

Kannan, R., Woo, H., Aggarwal, C. C., & Park, H. (2017b). Outlier detection for text data. *Proceedings of the 17th SIAM International Conference on Data Mining, SDM 2017*, 489–497. https://doi.org/10.1137/1.9781611974973.55

Kokatnoor, S. A., & Krishnan, B. (2020). Self-supervised learning based anomaly detection in online social media. *International Journal of Intelligent Engineering and Systems*, *13*(3), 446–456. https://doi.org/10.22266/IJIES2020.0630.40

Koppel, M., & Seidman, S. (2018). Detecting pseudepigraphic texts using novel similarity measures. *Digital Scholarship in the Humanities*, *33*(1), 72–81. https://doi.org/10.1093/llc/fqx011

Krommyda, M., Rigos, A., Bouklas, K., & Amditis, A. (2021). An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media. *Informatics*, *8*(1). https://doi.org/10.3390/informatics8010019

Kumar, S., & Bhatia, K. K. (2020). Semantic similarity and text summarization based novelty detection. *SN Applied Sciences*, *January*. https://doi.org/10.1007/s42452-020-2082-z

Kumari, R., Ashok, N., Ghosal, T., & Ekbal, A. (2021). Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. *Information Processing and Management*, *58*(5), 102631. https://doi.org/10.1016/j.ipm.2021.102631

Lazhar, F. (2018). Fuzzy clustering-based semi-supervised approach for outlier detection in big text data. *Progress in Artificial Intelligence*, *8*(1), 123–132. https://doi.org/10.1007/s13748-018-0165-5

Lee, D., Hyun, D., Han, J., & Yu, H. (2021). Out-of-Category Document Identification Using Target-Category Names as Weak Supervision. *Proceedings - IEEE International Conference on Data Mining, ICDM*, *2021-Decem*, 1186–1191. https://doi.org/10.1109/ICDM51629.2021.00041

Lochter, J. V., Pires, P. R., Bossolani, C., Yamakami, A., & Almeida, T. A. (2018). Evaluating the impact of corpora used to train distributed text representation models for noisy and short texts. *Proceedings of the International Joint Conference on Neural Networks*, *2018-July*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489355

Manolache, A., Brad, F., & Burceanu, E. (2021). *DATE: Detecting Anomalies in Text via Self-Supervision of Transformers*. 267–277. https://doi.org/10.18653/v1/2021.naacl-main.25

Mei, M., Guo, X., Williams, B. C., Doboli, S., Kenworthy, J. B., Paulus, P. B., & Minai, A. A. (2018). Using Semantic Clustering and Autoencoders for Detecting Novelty in Corpora of Short Texts. *Proceedings of the International Joint Conference on Neural Networks*, *2018-July*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489431

Mohaghegh, M., & Abdurakhmanov, A. (2021). Anomaly Detection in Text Data Sets using Character-Level Representation. *Journal of Physics: Conference Series*, *1880*(1), 1–6. https://doi.org/10.1088/1742-6596/1880/1/012028

Mohotti, W. A., & Nayak, R. (2020). Efficient Outlier Detection in Text Corpus Using Rare Frequency and Ranking. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *14*(6), 1–30.

Mu, J., Zhang, X., Li, Y., & Guo, J. (2021). Deep neural network for text anomaly detection in SIoT. *Computer Communications*, *178*(July), 286–296. https://doi.org/10.1016/j.comcom.2021.08.016

Muppudathi, S. S., & Krishnasamy, V. (2023). Anomaly Detection in Social Media Texts Using Optimal Convolutional Neural Network. *Intelligent Automation and Soft Computing*, *36*(1), 1027–1042. https://doi.org/10.32604/iasc.2023.031165

Nguyen, V. Q., Anh, T. N., & Yang, H. J. (2019). Real-time event detection using recurrent neural network in social sensors. *International Journal of Distributed Sensor Networks*, *15*(6). https://doi.org/10.1177/1550147719856492

Page, M. J., McKenzie, J. E., Bossuyt, P., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *Medicina Fluminensis*, *57*(4), 444–465. https://doi.org/10.21860/medflum2021_264903

Pizurica, N., & Tomovic, S. (2020). An approach for outlier and novelty detection for text data based on classifier confidence. *AI Communications*, *33*(3–6), 139–153. https://doi.org/10.3233/AIC-200649

Pv, S., & Bhanu, S. M. S. (2020). UbCadet: detection of compromised accounts in twitter based on user behavioural profiling. *Multimedia Tools and Applications*, *79*(27–28), 19349–19385. https://doi.org/10.1007/s11042-020-08721-z

Ray, P., Ganguli, B., & Chakrabarti, A. (2021). A Hybrid Approach of Bayesian Structural Time Series with LSTM to Identify the Influence of News Sentiment on Short-Term Forecasting of Stock Price. *IEEE Transactions on Computational Social Systems*, *8*(5), 1153–1162. https://doi.org/10.1109/TCSS.2021.3073964

Rose, R. L., Puranik, T. G., & Mavris, D. N. (2020). Natural language processing based method for clustering and analysis of aviation safety narratives. *Aerospace*, *7*(10), 1–22. https://doi.org/10.3390/aerospace7100143

Ruff, L., Vandermeulen, R., Schnake, T., & Kloft, M. (2019). Self-Attentive, Multi-Context One-Class Classification for Unsupervised Anomaly Detection on Text. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4061–4071.

Santiso, S., Pérez, A., Casillas, A., & Oronoz, M. (2020). Neural negated entity recognition in Spanish electronic health records. *Journal of Biomedical Informatics*, *105*(December 2019), 103419.

https://doi.org/10.1016/j.jbi.2020.103419

Sarna, G., & Bhatia, M. P. S. (2020). Structure-based analysis of different categories of cyberbullying in dynamic social network. *International Journal of Information Security and Privacy*, *14*(3), 1–17. https://doi.org/10.4018/IJISP.2020070101

Schmitt, M. F. L., & Spinosa, E. J. (2018). Outlier Detection on Semantic Space for Sentiment Analysis with Convolutional Neural Networks. *Proceedings of the International Joint Conference on Neural Networks*, *2018-July*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489200

Schulte, J. P., Giuntini, F. T., Nobre, R. A., Do Nascimento, K. C., Meneguette, R. I., Li, W., Gonçalves, V. P., & Filho, G. P. R. (2022). ELINAC: Autoencoder Approach for Electronic Invoices Data Clustering. *Applied Sciences (Switzerland)*, *12*(6). https://doi.org/10.3390/app12063008

Seki, K., Ikuta, Y., & Matsubayashi, Y. (2022). News-based business sentiment and its properties as an economic index. *Information Processing and Management*, *59*(2), 102795. https://doi.org/10.1016/j.ipm.2021.102795

Sepúlveda-Torres, R., Vicente, M., Saquete, E., Lloret, E., & Palomar, M. (2021). HeadlineStanceChecker: Exploiting summarization to detect headline disinformation. *Journal of Web Semantics*, *71*, 100660. https://doi.org/10.1016/j.websem.2021.100660

Serra, E., Akella, H., & Cuzzocrea, A. (2018). A Crowdsourcing Semi-Supervised LSTM Training Approach to Identify Novel Items in Emerging Artificial Intelligent Environments. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, 1479–1485. https://doi.org/10.1109/ICMLA.2018.00241

Shaalan, Y., Zhang, X., Chan, J., & Salehi, M. (2021). Detecting singleton spams in reviews via learning deep anomalous temporal aspect-sentiment patterns. In *Data Mining and Knowledge Discovery* (Vol. 35, Issue 2). Springer US. https://doi.org/10.1007/s10618-020-00725-5

Shi, X., Cai, L., & Song, H. (2019). Discovering potential technology opportunities for fuel cell vehicle firms: A multi-level patent portfolio-based approach. *Sustainability (Switzerland)*, *11*(22). https://doi.org/10.3390/su11226381

Sivakumar, C., Sathyanarayanan, D., Karthikeyan, P., & Velliangiri, S. (2022). An Improvised Method for Anomaly Detection in social media using Deep Learning. *Proceedings of the International Conference on Electronics and Renewable Systems, ICEARS 2022*, *Icears*, 1196–1200. https://doi.org/10.1109/ICEARS53579.2022.9751851

Sudha, K., & Suguna, N. (2019). Anomaly analysis based on meta-subspace approach for sentiment classification. *Journal of Intelligent and Fuzzy Systems*, *36*(4), 3403–3412. https://doi.org/10.3233/JIFS-181138

Suyanto, S., Meliana, S., Wahyuningrum, T., & Khomsah, S. (2022). A new nearest neighbor-based framework for diabetes detection. *Expert Systems with Applications*, *199*(March), 116857. https://doi.org/10.1016/j.eswa.2022.116857

Thomas, J. G., Mudur, S. P., & Shiri, N. (2019). Detecting anomalous behaviour from textual content in financial records. *Proceedings - 2019 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2019*, *i*, 373–377. https://doi.org/10.1145/3350546.3352550

Walkowiak, T., Datko, S., & Maciejewski, H. (2018). Algorithm based on modified angle-based outlier factor for open-set classification of text documents. *Applied Stochastic Models in Business and Industry*,

34(5), 718–729. https://doi.org/10.1002/asmb.2388

Wallace, D., & Kecahdi, T. (2019). Outlier Detection in Health Record Free-Text using Deep Learning. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 550–555. https://doi.org/10.1109/EMBC.2019.8857491

Wang, H. C., Chi, Y. C., & Hsin, P. L. (2018). Constructing patent maps using text mining to sustainably detect potential technological opportunities. *Sustainability (Switzerland)*, *10*(10), 1–18. https://doi.org/10.3390/su10103729

Wang, J., & Chen, Y. J. (2017). Technological Opportunity Analysis for the Telehealth Industry. *Proceedings - 2017 6th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2017*, 35–40. https://doi.org/10.1109/IIAI-AAI.2017.97

Wang, J., Kan, H., Meng, F., Mu, Q., Shi, G., & Xiao, X. (2020). Fake review detection based on multiple feature fusion and rolling collaborative training. *IEEE Access*, *8*, 182625–182639. https://doi.org/10.1109/ACCESS.2020.3028588

Wang, Z., Ma, L., & Zhang, Y. (2017). A hybrid machine learning method for finding depression related publications by eliminating outlier publications. *Proceedings - 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, *2017-Janua*, 171–176. https://doi.org/10.1109/IRI.2017.75

Yu, W., Boenninghoff, B., & Kolossa, D. (2021). Hybrid Representation Fusion for Twitter Hate Speech Identification. *CEUR Workshop Proceedings*, *3159*, 319–329.

Zhao, Q., Liu, J., Sullivan, N., Chang, K. C., Spina, J., Blasch, E., & Chen, G. (2021). *Anomaly detection of unstructured big data via semantic analysis and dynamic knowledge graph construction*. *April 2021*, 21. https://doi.org/10.1117/12.2589047

Zhu, J. W., & Zhang, Y. H. (2017). Chinese web short text subject clustering based on similarity upper approximation. *2017 International Conference on Computer Systems, Electronics and Control, ICCSEC 2017*, 1307–1310. https://doi.org/10.1109/ICCSEC.2017.8446684

Zhu, M., Aggarwal, C. C., Ma, S., Zhang, H., & Huai, J. (2017). Outlier detection in sparse data with factorization machines. *International Conference on Information and Knowledge Management, Proceedings*, *Part F1318*, 817–826. https://doi.org/10.1145/3132847.3132987