Journal of Computing in Higher Education (2021) 33:779–806
https://doi.org/10.1007/s12528-021-09295-x

IoT text analytics in smart education and beyond

Abdul Hanan Khan Mohammed[1] · Hrag-Harout Jebamikyous[1] · Dina Nawara[1] · Rasha Kashef[1]

Accepted: 18 August 2021 / Published online: 31 August 2021

Abstract

Data Analytics has become an essential part of the Internet of Things (IoT), mainly text analytics-related applications, since they can be utilized to benefit educational institutions, consumers, and enterprises. Text Analytics is excessively used in Smart Education after the emerging technologies such as personal computers, tablets, and even smartphones transformed the education system and improved the teaching methods by helping the teachers to evaluate the students' performance or determine the degree of similarity between a lecturer's and the students' posts in the discussion forum, and by collecting the students' feedback on the teaching method, in order to categorize each feedback into positive or negative, which will help the lecturers in optimizing their way of teaching. In this paper, we highlight the main components of IoT analytics, along with a comprehensive background of text analytics used techniques and applications. This paper provides a comprehensive survey and comparison of the leveraged IoT Text Analytics models and methods in Smart Education and many other applications.

Keywords IoT · Text mining · Big data · Sentiment analysis · Data analytics · Smart education

✉ Rasha Kashef
rkashef@ryerson.ca

Abdul Hanan Khan Mohammed
abdul1.mohammed@ryerson.ca

Hrag-Harout Jebamikyous
hjebamikyous@ryerson.ca

Dina Nawara
dina.nawara@ryerson.ca

[1] Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada

Introduction

IoT (Internet of Things) has become very popular and has been involved in many industries such as (E-Commerce, eHealth, smart transportation, smart home, etc.), which led to an increase in the number of connected devices and generated data. IoT has been widely adopted to enhance industrial efficiency, performance, communication, and profitability (Allama & Dhunny, 2019). IoT devices (sensors, actuators) collect and transfer massive, big, and versatile data. That introduced the urge to have IoT data analytics. Analytics is about deriving knowledge from data, such as revealing patterns, trends, and correlations, to enable concerned entities like (Industries, education, etc.) to make proper decisions based on the findings they get from analyzing these data using statistics and machine learning methods (Marjani, et al., 2016). One significant aspect of extracting information from unstructured data is text mining (Maheswari & Sathiaseelan, 2015). Figure 1 shows a schematic diagram for IoT devices and Text Mining. Unlike humans, machines can't easily understand the text straightforwardly; the text's interpretation is always accompanied by context, background, and different people's mutual interactions. Text mining plays a crucial role in many applications. For example, in social media like (Facebook, Twitter), they can understand the users' behaviors and identify their potential customers using Text Mining techniques (Dastanwala & Patel, 2016). Clustering and classification techniques can be applied to text to retrieve different information that can be utilized later in the application-based analysis. Text Mining has several applied methods (Information Retrieval, Information Extraction, Categorization, Natural Language Processing, Clustering).

Smart Education aims to solve several challenges that education with the conventional methods face, for example, geographical location, availability, accessibility of the classes, and teachers'-students' performance evaluation. Smart Education can be beneficial in dealing with knowledge and idea amongst students and lecturers. Such knowledge can solve various problems, which requires professional knowledge to be accurately resolved by learning related knowledge when making

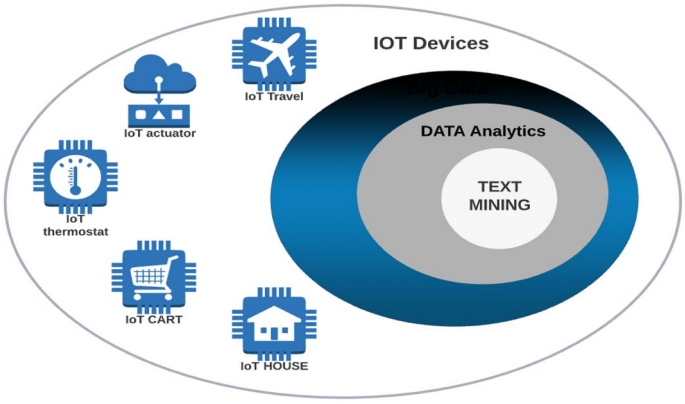

Fig. 1 IoT Data Analytics and Text Mining

decisions. Because of the emerging methods concerned with developing the current educational process, there is an increase in the amount of text-based data that can be analyzed to drive valuable insights. Text mining techniques utilize all text-based interactions such as (emails, discussion forums, or social media interactions), where these interactions can be explored using different text mining techniques as proposed and studied in the literature. Some of these techniques are referred to as (i) Sentiment analysis, (ii) Opinion mining, or (iii) Emotion analysis. Using these techniques leads to valuable insights into better decision-making and an enhanced student engagement process. For instance, opinion mining in smart learning is used as an application for evaluating the students' opinions about some courses provided in the learning process, therefore allowing the teachers to understand and act upon the needs of their students. However, sentiment analysis in smart education can be challenging to deploy due to the semantical variability of the used language. In addition to smart education, IoT text mining has received a broader spectrum of research opportunities in various other industrial fields as healthcare, e-commerce, smart transportation, and smart systems. For example, the authors conducted a literature review in (Ittoo et al., 2016) to review some famous state-of-art techniques for text analytics in the industry. Another survey was done in (Nair et al., 2020) focusing on the used text analytics techniques and algorithms and how they can be incorporated in domains such as e-health.

This paper focuses on the use cases proposed and studied by researchers in the literature incorporating different techniques and algorithms to enhance IoT smart education using text mining approaches. In this paper, we provide a road map for future research through our comprehensive survey on the current state-of-the-art IoT models using text mining in smart education and beyond. This paper has provided a comparative overview of the recent research regarding datasets used, modeling techniques, evaluation methods, and limitations. This paper is structured as: Sect. 2 presents a background to IoT, Big Data, and data analytics in IoT. Section 3 sheds light on text mining and its use cases. Section 4 introduces related work on IoT text mining in smart education. Section 5 discusses the various business applications of IoT text mining. Section 6 concludes the paper with the primary takeaways points. Finally, Sect. 7 shows the future directions of the current research.

Background on IoT, big data, and analytics

Internet of Things (IoT)

The Internet of Things *(IoT)* model has grown into technology for creating smart environments for human life that make life very much simpler for human interaction. Since this technology is strongly attached to real-world data, key issues arise in everyday businesses, such as data privacy and security in any IoT model. The Internet-of-Things (IoT) is concerned with building a smart environment to improve everyday life through uninterrupted connectivity. This is accomplished with sensors and devices' interconnectivity to facilitate assessments made via analysis of existing user data. The IoT technologies are used to interconnect

human beings and machines, whereby sensors and networks allow all participating nodes to communicate directly to share important data.

IoT devices framework

The design of an IoT architecture is categorized into five layers. Each layer has its responsibilities**,** where the physical layer collects the data and sends it to the network layer, which is responsible for the transmission of data.

Management and utilization of data are covered in the application layer. Therefore, an IoT device can adapt to any smart environment according to the user's needs. Each layer plays a dynamic role in the framework of an IoT device or IoT sensors. Below stated the functionality per layer.

- *Physical Layer* Physical objects and sensors are the hardware associated with the Perception layer. This layer provides data identification, data collection, and then data is processed. All this data is then sent to the next layer, which is the network layer.
- *Network Layer* In the network layer, the information is transmitted over secured lines using wired or wireless communication systems from IoT devices or sensors to the processing system. The network layer's output is then sent to the middleware layer, where decision-making or analytics are performed.
- *Middleware Layer* Management of services between the IoT devices is carried out in the Middleware layer. It also collects and stores the data from the network layer to perform decision-making or data analytics.
- *Application Layer* Different data implemented on IoT applications in smart industries, buildings, hospitals, universities, and health is accountable in the Application layer for global management.
- *Business Layer* Service and global management over IoT devices are held accountable in the Business layer. It also produces a business model, flowcharts, and graph based on the application layer's data and provides the results of processed operations.

IoT applications are different from other IT-related techniques due to their universal and embedded features that are essentially used in our daily lives. IoT architecture can be tagged as undefined or multi-defined as it involves various domains and technologies such as sensors, interactive devices. These interconnected technologies can then share data and communicate with each other and the outside world through dedicated APIs and gateways. These enable the data from these connected sensors and devices to be captured analytically. The IoT paradigm enables new research potentials for using data collected from these interconnected small devices (sensors and actuators). At the same time, it also introduces new challenges for the data analytics involved with using this large amount of data. Figure 2 demonstrates the different layers of IoT architecture.

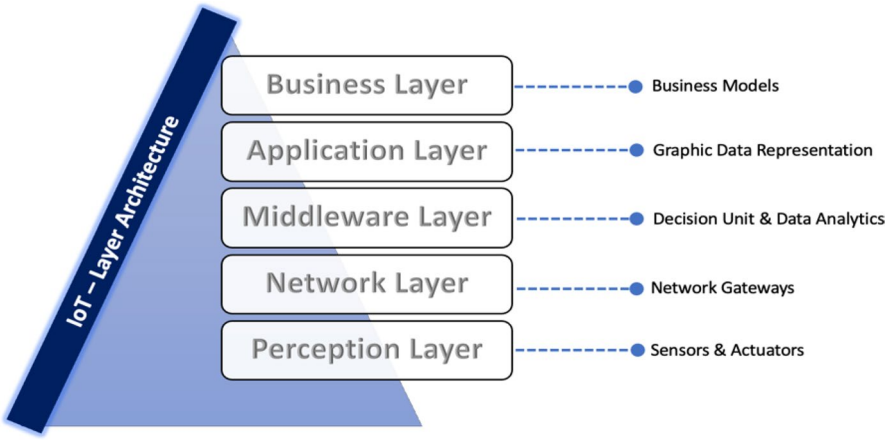

Fig. 2 The IoT architecture layers

Big data

In the last few years, the data volume has dramatically increased, and it is expected to grow more. According to Statista, the volume is expected to be 149 Zettabytes by 2024 (Statista, n.d.), as shown in Fig. 3. Data is classified into Structured, Semi-Structured, and Unstructured data. Big Data has different characteristics such as velocity, volume, veracity, value, and variety, known as (5 V's).

- *Volume* indicates that the size of the data is bigger than the traditional operational database data. Big Data can exceed Petabytes and Exabytes units (10^{15} bytes) and (10^{18} bytes), respectively.
- *Velocity* indicates a high data streaming rate, such as (video streams for traffic cameras). With the velocity comes the importance of the term "Data aging," which means data value against time (Osman, 2019).
- *Variety* indicates the different data formats, such as the data coming from text files, images, etc. The data can be classed as a structured, semi-structured, stream, or unstructured data.

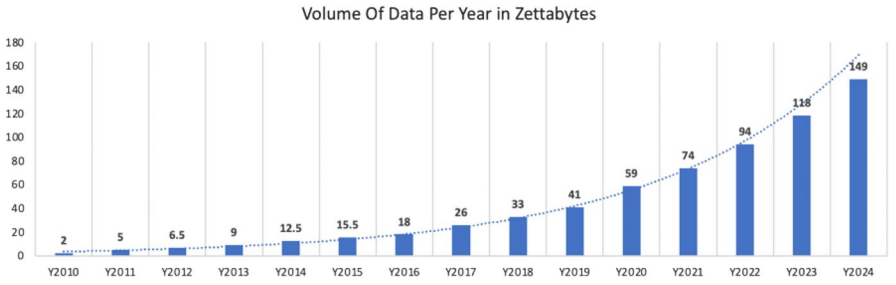

Fig. 3 The volume of data per year (Statista, n.d.)

- *Veracity* refers to how extent to which the data can be trusted. It is essential to have a reliable data source since false data will lead to inaccurate and misleading results.
- *Value* refers to the gained value of using this data. This measure comes as an assessment of whether or not the transformation is useful.

IoT-Analytics

Analytics increase the performance and efficiency of systems and industries by driving insightful knowledge from the raw data. IoT analytics are involved in major sectors such as operations and maintenance, supply chain, customer experience, smart cities, smart transportation, etc. (Patel et al., 2017). Data analytics in IoT can be categorized as descriptive, predictive, prescriptive, and diagnostic analytics, as shown in Fig. 4. Below we briefly define each category.

- *Descriptive analytics* works on raw data and helps us describe, summarize and visualize the IoT data. It can also be used as an initial step for further analysis.
- *Diagnostic analytics* understands the root causes and details behind why things have happened that way. It also utilizes data visualization.
- *Predictive analytics* which answers the question of "what is most likely going to happen?" forecasts and predicts future outcomes based on historical data. It contains machine learning and statistical models.
- *Prescriptive analytics* allows understanding the effect of each decision, as well as the connection between different decisions. It exploits future opportunities and minimizes the risks.

Text analytics in IoT

Text Mining refers to the process of identifying valuable information from structured or unstructured textual data. The methods adopted in this process range with different purposes and applicable to other document types such as current trend analysis, opinion mining, feedback analysis, and understanding specific research fields

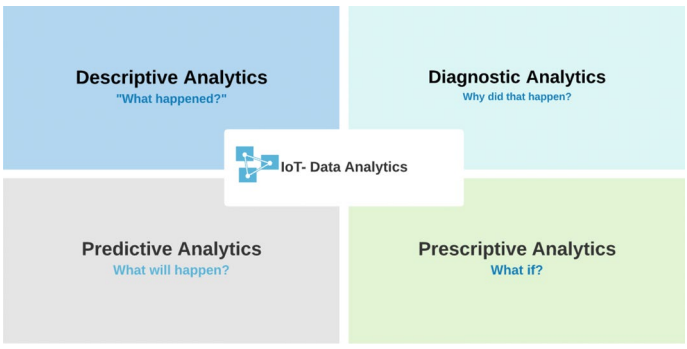

Fig. 4 IoT-Analytics Classifications

with scientific documents. Our research indicated an increasing trend of applying machine learning algorithms to textual data to discover hidden information from data. For example, identifying anonymous categories of documents using a clustering algorithm, identifying essential words, and automated spam emails classification using a classifier. Our research identified that visualization is critical for interpreting textual data as text mining incorporates various features and requires semantic interpretation.

Text mining tasks

The textual data has been involved in our day-to-day lives. For example, it is found in patients' records, news, etc. That led to a dramatic increase in the volume of text information. Some algorithms are deployed to find patterns from the text, and that process is called text analytics. Text analytics covers and employs related topics like Natural Language Processing (NLP) and Information Retrieval (IR). NLP means the understanding of the natural language. Information Retrieval (IR) facilitates information access, i.e. (access to documents) to understand and analyze this information (Allahyari et al., 2017).

- *Natural Language Processing (NLP)* is a powerful toolbox used to manage and identify text and access knowledge from unstructured data. It is used to segment, parse, extract and analyze text data.
- *Information Retrieval (IR)* is about searching ***documents*** within a collection of text documents. It is used to facilitate access to these documents. For example, some recommendation systems (Fayyaz et al., 2020) based on text analysis are utilizing IR approaches
- *Information Extraction (IE)* extracts information from semi-structured or unstructured data with a prior constraint that the information to be extracted. It is involved in biomedical text mining-related applications.
- *Text summarization* creates short and concise versions with essential information of several documents and can be classified into abstractive and extractive summarization.

The text mining process

The text mining process (Fig. 5) consists of multiple steps such as tokenization, filtering, converting to lower case, lemmatization, stemming, modeling, and training.

- *Tokenization* is the process of breaking larger sentences, paragraphs, or sequences into small or individual pieces called tokens. The importance of using tokenization in text analysis helps to identify words that have a string of characters and split them into small tokens. We can use tokenization to count the number of words in the given texts or the frequency of the repeated words.
- *Filtering* removes some words called stop-words, which frequently appear in the text, yet they do not have significant content information. This step

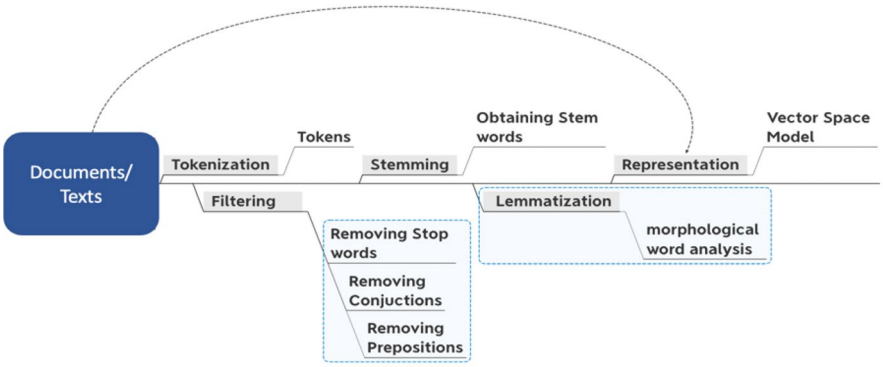

Fig. 5 The Text Mining Process Flow

gives the option to filter out the digits, punctuation, and words that do not add meaning to the body of the sentences. For example, consider the sentence "English is a universal language, where 'English', 'universal' and 'language' are the most significant words and 'is', 'an' are almost useless. In this way, we can filter out the useless words, punctuations, or any stop words.

- *Stemming* is the process of obtaining stem from words. It can also be referred to as the method of reducing a word to its stem or root. For example, let us consider a set of words, i.e., study, studied, and studying. These three words have different sentences but belong to the same root word, 'study'. After we stem the words, we will be left with only one word, i.e., study.
- *Lemmatization* is the process of considering morphological word analysis. In other words, it is a step-by-step process of acquiring the origin form of the word. It is similar to Stemming, but the root of the word in Lemmatization is referred to as lemma. The main advantage of using lemmatization is results of the reduced word are more accurate. For example, the word 'better' could be reduced to "bet" or "bett" if we have used the stemming process, but when the lemmatization process is used, the root of the word is reduced to the word 'good' which is more accurate.
- *Representation* in text mining, finding a suitable data model to represent documents and terms is crucial.

(1) The vector space model is commonly used to represent documents using the Term Frequency-Inverse Document Frequency (TF-IDF). TF-ID is a weighting factor used in Text Mining to measure a word's importance in a given document. Its value increased with the number of times a word appears in the document:

$$TFIDF_{ij} = TF_{ij} \times idf_i \tag{1}$$

Table 1 The one-hot encoding representation

Please	Write	Your	Homework
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Table 2 The bag-of-words representation

	Class	Documents
Training	–	It is unpredictable and lacks energy
Training	–	No jokes and very few laughs
Training	+	The future is yours
Test	?	Unpredictable with no future

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{2}$$

$$IDF_i = \log \frac{|D|}{\left|\left\{j|t_{it} \in d_j\right\}\right|} \tag{3}$$

(2) The N-grams language model is another well-known text representation technique that estimates the last word's probability given the previous words. N-gram is a sequence of N items collected from a sample of text or a speech corpus. If we consider the following sentence: "Please write your homework". when N = 2, it is a two-word sequence of words, such as "Please write", "write your", or "your homework". When N = 3, it is a three-word sequence of words, such as "Please write your", or "write your homework". For example, given the sequence of words "Please write your", the likelihood of the next word being "homework" is higher than "pen" or "pencil".

(3) The One-Hot Encoding is used for representing documents with a small vocabulary. If a document has a vocabulary with four words, it can be represented by 4-dimensional vectors. To represent a given word, we set the component of that vector to 1 and the rest of the words to 0, as shown in Table 1.

(4) The Bag-of-Words model is used when categorizing (classifying) text based on sentiment or verifying whether the text is spam. The text is represented as a bag of words, ignoring the order of the words and their original position in the text and only considering the frequency of the words. This method will be explained in the following sentiment analysis example of binary classification: positive (+) and negative (-). The model is trained on three sentences with different classes, and the model is built to classify the fourth sentence as either positive (+) or (−). In this example (shown in Table 2), the probability of the positive class is 1/3, and the probability of the negative class is 2/3. Hence, the probability of the

words “Unpredictable”, “with”, “no”, and “future” given the negative class is higher than the sample probability of the positive class. Therefore, the sentence “Unpredictable with no future” is classified as negative based on the training dataset.

Sentiment analysis and opinion mining

The amount of produced text data increases and expands dramatically, hence the need for effective algorithms and techniques to discover useful patterns and information of the text (Allahyari et al., 2017). As a result of using text mining techniques, there are applications as Sentiment analysis or opinion mining widely adopted for many text-related applications. Sentiment Analysis is an application of NLP (Natural Language Processing) and is also called opinion mining or emotional analysis. Sentiment analysis helps users with their decision-making process since it is involved in main domains such as E-commerce (e.g., eBay, Amazon), and social media (e.g., Facebook) to extract feedback on a specific product or service. Generally, sentiment analysis can be performed using supervised and unsupervised learning (Abirami, 2016) as shown in Fig. 6, such as (Support vector machines, Neural networks, Naïve Bayes, clustering, etc.). Sentiment analysis can be performed on (i) document level, (ii) sentence level, and (iii) feature level. Different approaches are used in sentiment analysis; the commonly used ones are machine learning and lexicon-based approaches.

(1) *Machine Learning Approaches*

Machine learning approaches are commonly used to generate sentiment classification models in the sentiment analysis domain. Whereas these models first label the training data by sentiment, a set of features is extracted from that training data to be used into the classification algorithm, such as Support Vector Machine, Naïve Bayes, Decision Trees, Neural Networks, etc. For example,

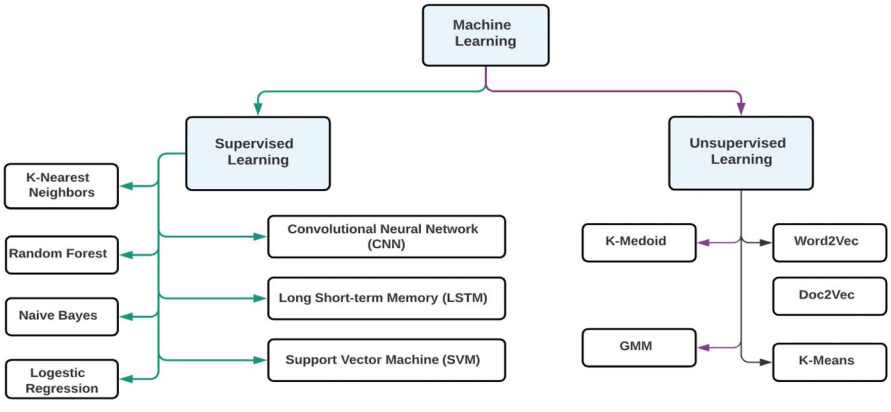

Fig. 6 A Taxonomy of the Various Machine Learning models

the Support Vector Machines and Decision tree-based sentiment classification model were introduced in (Rathi et al., 2018) to increase the tweets' sentiment classification accuracy. A (Robinson et al., 2016) machine learning prediction model was developed that learns from unstructured text to predict which students will complete the online course. The authors incorporated a lasso-regularized logistic regression model that balances variance and bias to make accurate predictions. A linear regression model was used to explore the relationship between students' participation and their learning gains using content analysis in the MOOC (Wanget al., 2015).

(2) *Lexicon Based Approaches*

In Lexicon-based approaches, the text is classified as positive, neutral, or negative based on the words contained in it. Unlike machine learning approaches, a Lexicon-based approach does not require a training phase and can be categorized into two categories: corpus-based and dictionary-based methods. Where corpus-based approach tries to capture co-occurring work patterns and hence it can determine the text's polarity. On the other hand, the dictionary-based approach utilizes antonyms, synonyms, and hierarchies in the lexical database (Wang et al., 2018). For example, an emotional dictionary of educational news using Chinese text was developed. The research work introduced in (Trinh et al., 2017) proposes a sentiment analysis framework based on combining learning-based and lexicon-based techniques for product feedback in the Vietnamese language. Sentiment analysis has a significant impact on education as it can observe the students' performance and learning patterns. For example, a lexicon-based method is introduced to identify the students' positive or negative feelings towards online teaching (PraveenKumar, 2020).

IoT text analytics in smart education

This section reviews the research work in IoT text analytics, especially in smart learning. This section focuses on covering related work to three main coherent in smart education, including learning analytics, feedback/performance analysis, and recommendation systems for intelligent learning.

Learning Analytics: The emerging technologies transformed the learning delivery approach, where they improved the teaching methods, students' engagement and increased the teaching accessibility tools. Recently, students can connect to their classes through their smart devices from any location and anytime. For instance, many proposals and mobile applications were adopted to create a Learning Management System (LMS), supporting e-learning techniques (Villegas-Ch et al., 2020). The researchers in (Yu & Zheng, 2017) incorporated data mining and natural language processing methods to analyze the students' learning behavior in Massive Online Open Courses (MOOCs). The outcome of the analysis aimed to evaluate the students' homework, postings, and answer contents. The author's main focus was on tendency analysis of the students learning since they needed to extract the emotional

tendencies in the network learning framework, which will help them give an overall assessment for different types of courses, teachers, or students. They used both text classification to classify the content of comments, discussions, and replies, then deploy an automatic clustering for them.

Feedback/Performance Analysis: Creating tools for instructors to assess their students' performance has a great impact on the learning process. Al-Ashmoery et al. (Al-Ashmoery & Messoussi, 2015) implemented a new learning analytics system for LMS to help the teachers assess the students' performance and text mining in online communication. The authors used already existing methods to compute sentence similarity on the text messages in LMS. The used technique measures the semantic similarity between the texts exchanged during an online communication session to calculate the degree of coherence in a discussion. Moreover, Buenaño-Fernández et al. in (Buenaño-Fernández et al., 2018) used text mining to evaluate students' interactions in online learning environments. The dataset used for the research consisted of 2751 tweets and 950 emails. After preprocessing the data, they used text mining techniques such as calculating the frequency of terms, analyzing concordances and groupings in n-grams, and then fed the processed data into the SVM classification model for opinion mining resulting in 75% accuracy. Kingsley et al., 2020 (Kingsley, et al., 2020) proposed the Educational Process Data mining model (EPDM) for the evaluation of text data collected in student's evaluation of teaching (SET), which benefits the students and teacher's relationship for calculated decision making. The data collected helps understand the students' words and opinions towards their teacher by seeing the gender differences. The work shown by the authors is the latest idea of using text mining for future educational development to deliver a more strong analysis of student's opinion towards their teachers. The authors in (Gomede et al., 2018) presented a model for developing the students' knowledge profile by creating key performance indicators to monitor the objectives' achievement using text and data mining techniques. The model used by the authors is a random forest model for classification and prediction and text mining techniques by using a dataset obtained from a Brazilian private k-9 elementary school to predict students' performance and behavior and recommendation to the teachers. Understanding the students' feedback is a very crucial step to enhance the overall learning experience; that is what the authors in (Aung & Myo, 2017) tried to do, where they analyzed the students' text feedback using the lexicon-based approach—that analysis aimed to enhance the teaching performance. The authors created a database of English sentiment words as a source of words' polarities. Both close-ended questions and open-ended questions collected the students' feedback. They collected comments from the University of Computer studies in Mandalay as input; then, they calculated the polarity score for the students' feedback, where they could eventually categorize the feedback into positive, negative, strongly positive, and weakly negative. Sentiment analysis was incorporated many times in the literature, using different algorithms; for example, Rani et al., 2017 (Rani & Kumar, 2017) proposed a sentiment analysis system that evaluates the course satisfaction index and teacher performance. The developed system classifies sentiments into positive and negative. Data was collected from Coursera and also from the University of SRS. The authors used a lexicon that had 40 languages. The data collected from SRS was used

for the system evaluation and validation. Also, Nkomo et al., 2020 (Nkomo et al., 2020) conducted a study to understand the value of lecture recordings. They analyzed the discussions on social media, especially Facebook, related to the value of the recorded lecture. The authors utilized data obtained from a Student Union forum on Facebook and incorporated Google NLP API.

Recommendations Systems for intelligent learning: We can see the role of smart education's recommendation systems utilizing text analytics in (Murad et al., 2018) where Murad et al. performed text mining analysis in the log discussion forum of recommendation systems for online learning by determining the degree of similarity between a lecturer's and the students' posts in the discussion forum. Their work consisted of three stages: they selected five classes randomly from the forum data in the first stage. In the second stage, they performed text mining analysis on the lecturers' and the students' posts. In the third stage, they used lecture notes as datasets in the form of corpus to analyze the text validation. They performed the objective task using the "doc2v" algorithm with vectorization. They reached two conclusions based on their proposed method, and the first conclusion is that the proposed method achieved 49% similarity between the lecturers' posts and the students' posts, and the second conclusion was that the similarity of the discussion between the students and the lecturers has a significant influence on the performance of the students. Also, in the Online Learning domain, the researchers in (Murad et al., 2018) illustrated the necessity for a recommendation system for innovative learning by using text mining combined with IoT data analytics modeling on datasets extracted from discussion forums shared by lecturers and students. The authors explore the strong positive influence of having recommendations in online learning tools on student learning outcomes by indicating a high percentage of similarity in discussion between lecturers and students. Tables 3 and 4 show various research in smart education using IoT text Analytics along with their evaluation metrics and limitations.

Business adoption of IoT text mining

This section summarizes the state-of-the-art research work in IoT text analytics, focusing on e-health, e-commerce, smart transportation, and intelligent systems.

IoT text mining in E-health

IoT text analysis is broadly used in the e-health field to benefit patients and healthcare professionals on all levels. A high percentage of individuals use the internet to relate to their health conditions, seeking medical advice, or related diagnostics. Many researchers have proposed text analytics models to complete the prescription understanding, follow-up, medicine, and diagnostics-based recommendation chains, as shown in Table 5.

A smart-enabled medicine box with a camera for reading the prescription is proposed in (Rumi et al., 2019). It is connected to the network so the doctors can monitor their patients' medicine consumption. They aimed to help the patients who can

Table 3 IoT-Text Analytics in Smart Education

Paper	Approach	Used Algorithm	Dataset
(Yu & Zheng, 2017)	Text Mining for tendency Analysis of Students learning	Word2vec	MOOCs
(Aung & Myo, 2017)	Sentiment Analysis for students' comments	Lexicon based approach	Comments from University of computer studies
(Rani & Kumar, 2017)	Sentiment Analysis for students' comments	NRC emotion lexicon	Coursera Database
(Nkomo et al., 2020)	Sentiment Analysis using SNA	NLP	Student Union forum on Facebook
(Murad et al., 2018)	Text Mining to implement a system of recommendations	Doc2v algorithm with vectorization	Datasets extracted from discussion forums shared by lecturers and students
(Al-Ashmoery & Messoussi, 2015)	Sentence Semantic similarity measure	Path Length, Resnik Similarity, Lin Similarity, Jiang-Conrath Distance, Wu, and Palmer Measure	LMS Log Data
(Buenaño-Fernández et al., 2018)	Text Mining	SVM	E-mails and Tweets
(Kingsley, et al., 2020)	Sentiment analysis for students' comments	EPDM Model	Student Opinion Surveys
(Gomede et al., 2018)	Text Mining to implement a system of recommendations	Random Forest	Brazilian Private School

Table 4 Evaluation Metrics and Limitations

Paper	Evaluation	Faced Limitations
(Yu & Zheng, 2017)	N/A	The difficulty of Information extraction
(Aung & Myo, 2017)	Comparison between a lexicon-based method and the existing state of art techniques	N/A
(Rani & Kumar, 2017)	Different students' surveys analysis	Lack of Multilingual capabilities
(Nkomo et al., 2020)	N/A	Context of feedback is not included
(Murad et al., 2018)	The similarity between students' posts and courses	lack of used algorithms'' efficiency measures
(Buenaño-Fernández, Villegas-Ch et al., 2018)	N/A	Non-English Language understanding
(Kingsley, et al., 2020)	Sentiment Score measurements' comparisons	Model is customized on specific datasets
(Gomede et al., 2018)	Accuracy	Real-time analysis

Table 5 IoT-text analytics in E-health applications

Paper	Approach	Used algorithm	Dataset
(Rumi et al., 2019)	Extracting text from images	Maximally stable extremal regions	Simulation
(Meena & Bai, 2019a Study on Machine learning based Social Media and Sentiment analysis for medical data applications, 2019)	sentiment analysis on social media	Lexical and statistical analysis with tokenization and use of a bag of words	Social media platforms
(Pendyala & Figueira, 2017)	Information retrieval approach	K means for clustering	Discharge sheets
(Gupta et al., 2017)	If, else rules	N/A	Simulation
(Meena & Bai, 2019b Study on Machine learning based Social Media and Sentiment analysis for medical data applications, 2019)	Lexical analysis	statistical analysis	Social media platforms
(Raji et al., 2016), (Zaman & Mamun, 2017)	N/A	Naïve Bayes and SVM	Simulation
(Asthana & Megahed, 2017)	Text mining to analyze unstructured medical history to produce structured demographic attributes	Tf-idf based text mining algorithm. Decision Tree, Logistic Regression, LibSVM, and OneR	Publicly available electronic health records (HER)

make mistakes by taking the wrong medication. Their model is used to scan the provided prescription and retrieve information by applying image processing techniques and then using text area detection and the Maximally Stable External Regions (MSER) algorithm. The authors used Optical Character Recognition (OCR) for text recognition and evaluated their model using average accuracy. Sentiment analysis on the different social media platforms to find the users' perception regarding certain drugs or diseases' symptoms is introduced in (Meena & Bai, Study on Machine learning based Social Media and Sentiment analysis for medical data applications, 2019a). They extracted the tweets related to cancer and computed the number of likes on some drug responses. Mobile and cloud technologies were utilized to perform diagnosis using the vector support model in text mining (Pendyala & Figueira, 2017). They considered the diagnosis as a classification problem. They presented each set of symptoms as a vector and used cosine similarity in the implementation to relate similar diseases and symptoms, and they clustered the data using k-means. A health monitoring system for patients with Alzheimer's was proposed to provide monitoring that can be accessed anytime (Gupta et al., 2017). They also made use of mobile and cloud technologies and created different models according to the network. For example, in GSM, the transmitter transmits patients' information in the form of a text message. The authors used sensors such as (optical sensors, heart rate counter, conditional signaling unit, temperature sensor) and evaluated their model using overall patient satisfaction. The authors in (Meena & Bai, Study on Machine learning based Social Media and Sentiment analysis for medical data applications, 2019b) studied cancer disease using text analysis, precisely sentiment analysis. They involved google search, Twitter, and other platforms in their study as a source of healthcare data analysis. A real-time chronic disease monitoring system is presented and analyzed in (Raji et al., 2016). They relied on wearables sensors to measure temperature, heart rate, and blood pressure. These sensors monitored patients' vital signs and stored the data in the form of text. Without the nurse's help, the system turns the text data into status for the patients. The authors classified the signs using Naïve Bayes and SVM classifiers and evaluated their model using precision, recall, and F-measure. A study in (Zaman & Mamun, 2017) evaluates current healthcare systems to classify the different healthcare applications. They considered their study using a questionnaire. Table 5 provides a summary of various e-health applications using text analytics and IoT.

IoT text mining in smart transportation

Many sensors and mobile applications are being employed in transportation systems nowadays for a better and safer traveling experience. Also, individuals share their opinions and thoughts on their travel plans or travel experience using social media platforms in the form of text. Text analytics can benefit intelligent transportation companies and consumers; that is why some research has been conducted in that area. For example, in (Ali et al., 2019), the authors proposed a new fuzzy ontology-based approach to improve feature extraction efficiency in text, using the LSTM approach. Their method exports the features from social media text as

climate, roads, accidents, vehicles, and applied feature polarity identification and evaluated their model using Precision, Recall, and Accuracy. Urban public services problems such as traffic and security were identified using text analytics (Gonzalez et al., 2020). They used the "Tweepy" library to search for tweets and applied it on social media platforms, such as Twitter. The authors presented their findings in a dashboard to indicate any security or traffic issues related to any selected area. The impact of the tweets when using Uber services is measured in (Ulloa et al., 2016). They used social media APIs for collecting the data and performed an information extraction approach based on keyword matching. The authors applied a sequential classification approach to detect the mentions. They used a lexicon-based algorithm to classify the sentiment category (neutral, positive, or negative).

In (Chaturvedi et al., 2020), the authors proposed a framework to detect traffic detection using word embedding and different machine learning models using Twitter. The proposed approach was divided into four phases, i.e., in the first phase, they have gathered geo-tagged tweets based on traffic-related keywords and non-traffic keywords. These keywords are split using the Natural Language toolkit library. The techniques to detect traffic patterns using Twitter which has information related to traffic incidents, were applied using Bag-of-words, TF-IDF, and Word2Vec with different machine learning algorithms such as KNN, Naïve Bayes, SVM, and a neural network model. An NLP-based back propagation neural network for predicting passengers in public transportation using Smart card data is presented in (Dou et al., 2015). This study used a similar concept of a "bag of words" for building the user profile. The back-propagation neural network approach was used on the results of clustering to generate passenger predictions. Table 6 summarizes the proposed techniques in the Smart transportation domain.

IoT Text Mining in E-Commerce

Text analytics has been widely used in the E-Commerce sector after the expansion of E-Commerce in the past decade, with expected E-Commerce sales to reach $4.2 Trillion and $6.5 Trillion by the end of 2020 2023, respectively (Statista, n.d.). As shown in Fig. 7.

Text mining is primarily used to mine and summarize the customers' reviews and comments on consumer products, which helps E-Commerce based businesses identify frequently occurred issues and enhance the quality of the products and services. In (Rangu et al., 2017), the author's goal was to build different text mining techniques to identify and classify the customers' complaints about a product. To extract the data from community sites, they used a chrome extension API named "API–KIMONO." To prepare the API, they have chosen six fields for each comment, "Subject": gives a brief info about the comment, "Username": shows the name of the user who wrote the comment, "User Type": There are many types of users, such as Contributor, Explorer, Member, New, etc., "Comment": it is the actual user's review about a product, "No. of Likes": gives the number of users that liked a given comment, "Date Time": gives the timestamp value of a given comment. In the preprocessing step, they used an R package called "tm," which contains 174

Table 6 IoT-Text Analytics in Smart Transportations

Reference	Approach	Used algorithm	Dataset
(Ali et al., 2019)	Feature extraction, sentiment classification	CNN, LSTM	social media platforms 500,000 sentences
(Gonzalez et al., 2020)	sentiment classification	N/A	Social media platforms
(Ulloa et al., 2016)	Information extraction	Lexicon based	Social media platforms
(Chaturvedi et al., 2020)	Bag-of-words, TF-IDF, Word2Vec	KNN, SVM, Naïve Bayes	Real-time publicly published tweets are collected using the Twitter streaming API
(Dou et al., 2015)	Bag-of-words	NLP based back propagation neural network	Public transport platforms

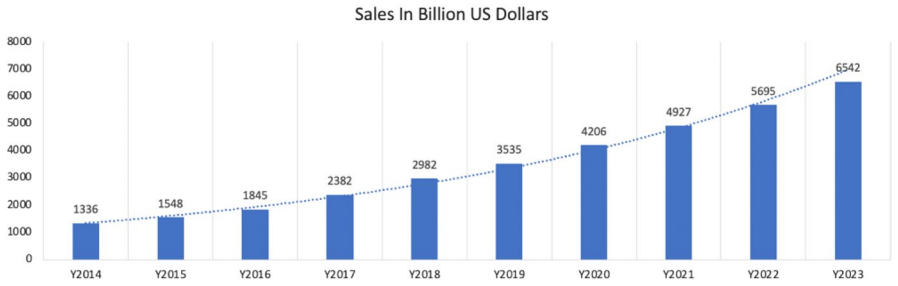

Fig. 7 The Retail E-Commerce sales worldwide 2014—2023

commonly used stop words in the English language and other domain-specific stop words that need to be removed because they do not add value to the customer review analysis. Even though Emoticons are a very useful way of expressing the customers' overall experience, they require a different method to analyze their meanings. Hence, the authors removed the reviews (few) with emoticons during preprocessing. Then they applied lemmatization and stemming to minimize the vocabulary size and created Term Document Matrix (TDM) to convert the processed data (text) to a structured format. They sorted the word list of every comment based on its frequency and chose every comment's top ten words to build the classifiers. After the preprocessing step, they tested a Support Vector Machine (SVM) classifier based on 8–12 words per comment, and they achieved the highest accuracy with comments that contained more than 10 words. They removed all the comments with less than 10 words and tested the data on the following three classification models:

- SVM: For the train/test split, they've used Stratified Random Sampling and used different SVM kernels. It achieved the highest classification accuracy of 69%.
- Bayes: Achieved classification accuracy of 57%.
- Random Forest: Achieved classification accuracy of 59%.

A Deep Learning (DL-based) E-Commerce sentiment classification model comprises four main steps (Zhang, 2020): Data Acquisition, Data Preprocessing, Emotion Annotation, and CNN Emotion Classification. In the Data Acquisition step, Octopus Collector's tool was used to crawl the customer comments regarding a tablet computer on a "Jingdong" platform. Due to the restrictions of the mentioned platform, they were able to crawl only 14,831 comments. In the Data Preprocessing step, duplicate comments and spam comments need to be removed from the dataset before labeling the dataset, leaving them with 14,630 comments (6392 positive comments, 3564 medium comments, 4674 negative comments). In the Dataset Labeling step, they performed data annotation based on sentiment lexicon, which consists of three modules: Segmentation of the preprocessed comments, identify emotional, degree, and negative words, and eventually calculate the emotional score to get the classification and label. The CNN sentiment classification step used "TensorFlow" as their back-end library and Keras as their front-end library to build their deep learning model. They split the data into 80% training set and 20% test set. They

extracted the required features from the training set to train the model and used the test set to predict the data class and model accuracy. They achieved a model accuracy of 0.8724. They then evaluated the model by comparing it with a Support Vector Machine (SVM) classifier, which resulted in 0.8297 accuracy. Their proposed method improved the SVM accuracy by 4%.

In (Rahardja et al., 2019), the author's objective was to analyze the customers' comments on an E-Commerce platform using the K-Medoid clustering algorithm. Their proposed architecture consists of two phases: Preprocessing phase and Clustering with the K-Medoid phase. In the Data Collection step, they retrieved 88 comments from two websites, namely, tokopedia.com and shopie.com. The Preprocessing step consisted of the following five stages (i) Tokenization (ii) Filtering stop words (iii) Transform cases (iv) Stemming (v) TF-IDF. In the Clustering step, they used the K-Medoid clustering algorithm, a partition-based unsupervised clustering algorithm like K-Means. But instead of clustering based on the means of the data, it uses actual data points to cluster the data. They evaluated the proposed model based on a two-step evaluation: Davies-Bouldin Index (DB) and Average in cluster distance.

Three models: Decision Tree, K-Nearest Neighbor (KNN), and Naïve Bayes to perform Sentiment Analysis on Tweets about Tokopedia and Bukalapak with 50 records each (Bayhaqy et al., 2018). They performed similar Preprocessing techniques as (Rahardja et al., 2019). They achieved the best results with the Naïve Bayes algorithm with an accuracy of 77%, precision of 88.50%, and recall of 64%. In (Yıldırım et al., 2019), the authors worked on a Real-World text classification application for an E-Commerce platform, and they collected the data from a fashion E-Commerce platform called "Morhipo", which consists of 1.2 Million rows and 24 columns (features) of data. The Preprocessing step consisted of 8 sub-steps: Convert uppercase letters to lowercase, Fix encoding errors, Removal of unnecessary characters, Translation of specific words from English to Turkish, Removal of punctuation marks, Removal of stop words, Stemming, and Vectorization. To achieve their goal, they built five classifiers: Linear Support Vector Classifier (LinearSVC), Stochastic Gradient Descent (SGD), Ridge Classifier (RC), Complement Naïve Bayes (CNB), Multinomial Naïve Bayes (MNB). Out of the mentioned five classifiers, LinearSVC outperformed all other classifiers with a Mean Accuracy of 96.08% and Standard Deviation (SD) of 5.65%. Table 7 presents a summary of the surveyed methods.

Table 7 IoT-Text analytics in E-commerce

Reference	Approach	Used Algorithm	Dataset
(Rangu et al., 2017)	Text Mining	SVM, Bayes, Random Forest	Community Cites
(Zhang, 2020)	Sentiment Classification	SVM, Deep Learning	E-Commerce Platform
(Rahardja et al., 2019)	Sentiment Analysis	K-Medoid	E-commerce Platforms
(Bayhaqy et al., 2018)	Sentiment Analysis	Decision Tree, KNN, Naïve Bayes	twitter
(Yıldırım et al., 2019)	Text Classification	LinearSVC, SGD, RC, CNB, MNB	E-Commerce Platform

IoT text mining in smart systems

The Internet of Things and Text Mining are combined to create new means of quality awareness, Artificial Intelligence, and support services. IoT is a technology that inputs from IoT sensors, people, online forums, maps, etc. These inputs are then used to aid information extraction in Smart Systems that enable users to create collective awareness, efficiencies, and better decision-making. With the advancements in these technologies, we live in times where objects are communicating with other objects over networks. These interactable objects range from personalized consumer appliances and wearables to big-scale IT infrastructures. Data generated from every connected source needs to be managed, analyzed, and trained with new information to take full advantage of the large amounts of undiscovered information to enable seamless interoperability of the connected objects. Smart Systems incorporate sensors, machines, and a wide range of devices producing massive amounts of structured and unstructured data. Retrieving previously undiscovered information from this large data source requires a new class of data modeling, management, and analytics tools. Applying IoT data analytics and text mining techniques to these data sources can provide numerous benefits like improved productivity, better detection of signal anomalies and operational risks, etc. In a perfectly connected world of intelligent systems, all the connected objects and machines will produce large amounts of valuable information. Text mining provides the ability to detect patterns in the data collected from these sources. A recent research in (Lim & Maglio, 2018) used text mining to develop a data-driven model to understand the smart system's framework. We can find smart systems everywhere, such as smart thermostats in homes and self-driving cars in the transportation, energy, and healthcare sectors. The authors of this research highlight the need for a comprehensive understanding of such systems in the literature. They used the dataset in simple text, and they analyzed and preprocessed the data using a combination of machine learning algorithms and metrics. The author in (Alzahrani, 2018) investigates the use of the Internet of things model for data analytics, concentrating on opinion mining and sentiment analysis. He presents integrating the Internet of Things domain with the computational linguistics domain by illustrating text mining as a pre-processing and cleansing step that influences the analysis's outcome. This pre-processing of texts included sentence and word tokenization, handling hyphens, punctuations, and numbers. A study in (Hong et al., 2018) highlights the importance of understanding the current state of IoT smart services and the role of customer thoughts in IoT for smart home services. Therefore, by using text mining analysis on a leading Korean telecommunication company, this study produces valuable IoT data analytics research results on the background of usage, functionality, and benefits of smart home IoT services by analyzing online customer reviews. Future sign searching techniques for smart grid IoT data analytics have been highlighted, focusing on text mining and its limitations (Park & Cho, 2017). The authors of this paper justify their usage of text mining to enable effective investigation of huge documents quickly. The authors also provide future research directions to overcome the limitations of this technique. A new smart recommendation system for personalized wearables and IoT smart solutions is offered in (Asthana & Megahed, 2017). They implemented the system using

text mining. They first identify the diseases a person may be at risk and perform data analytics algorithms on every individual's unstructured medical history, thereby producing structured data and then feeding it to a machine learning classification model that predicts eventual diseases. The authors illustrate the framework by mapping these recognized diseases to the attributes that need to be measured to monitor them. The developed framework then adopts a mathematical optimization model to recommend the individual's optimal wearable devices and IoT solutions. A recent study in (Xylogiannopoulos et al., 2017) targets text mining in unstructured and scrambled datasets such as smartphone electronic messages for digital forensics analytics. The authors highlight that the proposed framework in this paper applies nominal preprocessing by eradicating selected lexical characters to produce a continuous string of data and identify all the repeated patterns in the dataset using a modified of All Repeated Patterns Detection (ARPaD) algorithm and (LERP-RSA) Longest Expected Repeated Pattern Reduced Suffix Array data structure. Table 8 compares the different Smart systems proposed techniques.

Discussion and conclusion

IoT is considered the revolutionary transition in technology. Current trends in IoT technology provide extensive opportunities in big data analytics collected by IoT devices. Textual documents or forums are usually recovered from text analytics or data mining to discover unknown information and analyze the connectivity between documents. These mining techniques deliver effective predictive solutions for data analytics. Text Mining has shown great adoption in many applications, including smart education; however, we believe that these technologies are still in their early stages. Several prevailing research challenges have not yet been sought. Our survey paper compared the famous classification techniques used in text analytics, especially in the sentiment analysis domain, such as K-Nearest Neighbor, CNN, Naïve Bayes, and Support vector machine, where we compared their different applications' adoption. We covered three main coherent in smart education, including learning analytics, feedback/performance analysis, and recommendation systems for smart learning. Techniques such as natural language processing (NLP) and text mining contributed to providing the needed tools to enhance the learning experience. The researchers could utilize different classifiers to implement their models with satisfactory outcomes and accuracy measures. Support Vector Machine (SVM), and Random Forest (RF) algorithms were widely incorporated by various researchers in the smart education domain, as concluded in our literature. Also, different datasets were used in the literature, such as (tweets, courses' feedback, student surveys, questionnaires, etc.) due to their availability. The researchers can deploy more techniques aiming for continuous improvement. The smart health sector utilizes text analytics in various forms, such as monitoring prescription consumption of users through smart-camera enabled medicine box to track users' perception of new drugs on social media. Our survey shows that the smart transportation industry might benefit from the analyzed sectors by adopting text analytics, such as KNN, CNN, and LSTM, in their business model. These text analytics techniques could report the

Table 8 IoT-Text Analytics in Smart Systems

Paper	Approach	Used Algorithm	Dataset
(Hong, et al., 2018)	Text mining analysis	CONvergence of iterated CORrelations (CONCOR) analysis, followed by cross-analysis of the 20 most frequently used keywords	Online consumer reviews
(Park & Cho, 2017)	Lexical and Statistical Analysis	Keyword emergence map (KEM) and a keyword issue map (KIM) using the degree of visibility (DoV) and the degree of diffusion (DoD)	Radian6, a big data service, was used to retrieve textual data that can identify smart grid trends
(Lim & Maglio, 2018)	Text Mining	GMM based Clustering	Scientific Literature and News Articles
(Alzahrani, 2018)	Text mining, network analysis, and exploratory factor analysis	Combination of metrics and machine learning algorithms to preprocess and analyze text data	Google News platform service-based dataset
(Xylogiannopoulos et al., 2017)	Text mining in unstructured and scrambled datasets	Longest Expected Repeated Pattern Reduced Suffix Array (LERP-RSA) data structure and a variant of All Repeated Patterns Detection (ARPaD) algorithm	Smartphone electronic messages for digital forensics

real-time data regarding the climate, roads, accidents on a route through information retrieval of text extracted from social media. Besides the aforementioned main points, we conducted this survey to understand the comprehensive nature of IoT data with text mining processes focusing on smart education and other applications. We found a research gap in IoT and text analytics along with the adopted algorithms, especially for smart services. We discovered many limitations in the data gathering process for this survey, which supports our claim of the unavailability of adequate research in this field. This research paper was conducted to analyze the state-of-the-art techniques in smart education using IoT text mining to gauge the status of the research in this domain. This paper also compares various text mining techniques in other industrial applications as e-Commerce, Smart Transportation, eHealth, and smart service systems to assess the cross-platform applicability and scalability of the algorithms analyzed.

Future directions

IoT devices used in Smart Education, Smart Transportation, and E-health is an evolving concept where the practicality in universities and colleges are at low scale. The research in this area is minimal, and we were able to retrieve only a few papers that discussed the theoretical aspect of IoT-based smart services. Therefore, the future work in our research would focus on analyzing the challenges and limitations of IoT-based Smart Education to implement practical and efficient solutions. We would emphasize the comparison of the speed of connectivity between the devices and their communication efficiency based on the performance metrics and results of existing text analytics tools that have been implemented in Smart Education. We will also highlight the significance of the security risks involved in the data collection of the personal information of students and teachers and ways to improve the security of the resources to protect personal data. There is also a research gap in combining IoT and big text data analytics, which requires more research and will also be a focal point of our future directions.

References

Abirami, A. M., & Gayathri, V. (2017). A survey on sentiment analysis methods and approach. In *2016 Eighth International Conference on Advanced Computing (ICoAC)* (pp. 72-76). IEEE.

Al-Ashmoery, Y., & Messoussi, R. (2015). Learning analytics system for assessing students' performance quality and text mining in online communication. *Intelligent Systems and Computer Vision (ISCV).*

Ali, F., El-Sappagh, S., & Kwak, D. (2019). Fuzzy ontology and LSTM-based text mining: A transportation network monitoring system for assisting travel. *sensors*.

Allahyari, M., Pouriyeh, S., Assef, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief survey of text mining: Classification, clustering and extraction techniques. *Computation and Language.*

Allama, Z., & Dhunny, Z. A. (2019). On big data, artificial intelligence and smart cities. *Cities - Elsevier, 89*, 80–91.

Alzahrani, S. M. (2018). Development of IoT mining machine for Twitter sentiment analysis: Mining in the cloud and results on the mirror. *15th Learning and Technology Conference (L&T*, (pp. 86–95). Jeddah.

Asthana, S., & Megahed, A. (2017). A recommendation system for proactive health monitoring using IoT and wearable technologies. *IEEE International Conference on AI & Mobile Services (AIMS)*, (pp. 14–21). Honolulu.

Aung, K. Z., & Myo, N. N. (2017). Sentiment analysis of students' comment using lexicon based approach. *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS).*

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018). Sentiment analysis about E-commerce from tweets using decision tree, K-Nearest Neighbor, and Naïve bayes. *International Conference on Orange Technologies (ICOT).* Nusa Dua, BALI, Indonesia.

Buenaño-Fernández, D., Villegas-Ch, W., & Luján-Mora, S. (2018). Using text mining to evaluate student interaction in virtual learning environments. *IEEE World Engineering Education Conference (EDUNINE).*

Chaturvedi, N., Toshniwal, D., & Parida, M. (2020). Harnessing social interactions on twitter for smart transportation using machine learning. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 281-290). Springer, Cham.

Dastanwala, P. B., & Patel, V. (2016). A review on social audience identification on twitter using text mining methods. *IEEE WiSPNET.*

Dou, M., He, T., Yin, H., Zhou, X., Chen, Z., & Luo, B. (2015). Predicting passengers in public transportation using smart card data. In *Australasian Database Conference* (pp. 28-40). Springer, Cham

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Applied Science, 10*(21), 7748.

Gomede, E., Gaffo, F., Briganó, G., De Barros, R., & Mendes, L. (2018). Application of computational intelligence to improve education in smart cities. *Sensors, 18*(1), 267.

Gonzalez, M., Viana-Barrero, J., & Acosta-Vargas, P. (2020). Text mining in smart cities to identify Urban events and public service problems. *dvances in Artificial Intelligence, Software and Systems Engineering.*

Gupta, N., Saeed, H., Jha, S., Chahande, M., & Pandey, S. (2017). IoT based health monitoring systems. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).*

Hong, J., Suk, J., Hwang, H., Kim, D., Kim, K., & Jeong, Y. (2018). Text mining analysis of online consumer reviews on home IoT services.

Ittoo, A., Nguyen, L. M., & Bosch, A. v. (2016). Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry Elsavier, 78*.

Kingsley, O., Arturo, A.-P., Camacho-Zuñiga, C., Nisrine, H., Nakamura, E. L., & al., e. (2020). Impact of students evaluation of teaching: a text analysis of the teachers qualities by gender. *International Journal of Educational Technology in Higher Education, Heidelberg, 17*(1).

Lim, C., & Maglio, P. (2018). Data-driven understanding of smart service systems through text mining. *Service Science.*

Maheswari, M. U., & Sathiaseelan, D. J. (2015). Text mining: survey on techniques and applications. *International Journal of Science and Research (IJSR)*, 2319–7064.

Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A., Siddiqa, A., & Yaqoob, I. (2016). Big IoT data analytics: Architecture, opportunities, and open research challenges. *IEEE Access*.

Meena, R., & Bai, V. T. (2019). Study on Machine learning based Social Media and Sentiment analysis for medical data applications. *Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2019).*

Meena, R., & Bai, V. T. (2019). Study on Machine learning based Social Media and Sentiment analysis for medical data applications. *Proceedings of the Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2019).*

Murad, D. F., Heryadi, Y., Isa, S. M., Budiharto, W., & Wijanarko, B. D. (2018). Text Mining Analysis in the Log Discussion Forum for Online Learning Recommendation Systems. *2018 International Seminar on Research of Information Technology and Intelligent System.*

Nair, P. C., Gupta, D., & Devi, B. I. (2020). A Survey of text mining approaches, techniques, and tools on discharge summaries. In *Advances in Intelligent Systems and Computing book series (AISC, volume 1086)* (pp. 331–348). Springer.

Nkomo, L. M., Ndukwe, I. G., & Daniel, B. K. (2020). Social network and sentiment analysis: Investigation of students' perspectives on lecture recording. *IEEE Access , 8*.

Osman, A. M. (2019). A novel big data analytics framework for smart cities. *Future Generation Computer Systems Elsavier, 91*, 620–633.

Park, C., & Cho, S. (2017). Future sign detection in smart grids through text mining. *Energy Procedia*, (pp. 79–85).

Patel, P., Ali, M. I., & Sheth, A. (2017). On using the intelligent edge for iot analytics. *IEEE Intelligent Systems , 32*(5).

Pendyala, V. S., & Figueira, S. (2017). Automated medical diagnosis from clinical data. *IEEE Third International Conference on Big Data Computing Service and Applications.*

PraveenKumar, T. (2020). Exploring the students feelings and emotion towards online teaching: Sentimental analysis approach. *International Working Conference on Transfer and Diffusion of IT.*

Rahardja, U., Hariguna, T., & Baihaqi, W. M. (2019). Opinion mining on E-commerce data using sentiment analysis and K-Medoid clustering. *2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media).*

Raji, A., Jeyasheeli, P. G., & Jenitha, T. (2016). IoT Based classification of vital signs data for chronic disease monitoring. *10th International Conference on Intelligent Systems and Control (ISCO).*

Rangu, C., Chatterjee, S., & Valluru, S. (2017). Text mining approach for product quality enhancement: (Improving product quality through machine learning). *IEEE International Advance Computing Conference, IACC.*

Rani, S., & Kumar, P. (2017). A sentiment analysis system to improve teaching and learning. *Computer, 50*(5).

Rathi, M., Malik, A., Varshney, D., Sharma, R., & Mendiratta, S. (2018). Sentiment analysis of tweets using machine learning approach. *Proceedings of 2018 Eleventh International Conference on Contemporary Computing.*

Statista. (n.d.). Retrieved from https://www.statista.com/statistics/871513/worldwide-data-created/

Statista. (n.d.). Retrieved from https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016). forecasting student achievement in MOOCs with natural language processing. *ICPS Proceedings.*

Rumi, R. I., Pavel, M. I., Islam, E., Shakir, M. B., & Hossain, M. A. (2019). IoT enabled prescription reading smart medicine dispenser implementing maximally stable extremal regions and OCR. *Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2019).*

Trinh, S., Nguyen, L., & Vo, M. (2017). Combining lexicon-based and learning-based methods for sentiment analysis for product reviews in Vietnamese language. *International Conference on Computer and Information Science.*

Ulloa, D., Saleiro, P., Rossetti, R. J., & Silva, E. R. (2016). Mining social media for open innovation in transportation systems. *19th International Conference on Intelligent Transportation Systems (ITSC).*

Villegas-Ch, W., Román-Cañizares, M., & Palacios-Pacheco, X. (2020). Improvement of an online education model with the integration of machine learning and data analysis in an LMS. *Applied Sciences, 10*(15).

Wang, B., Gao, L., An, T., Meng, M., & Zhang, T. (2018). A method of educational news classification based on emotional dictionary. *2018 Chinese Control And Decision Conference (CCDC).*

Wang, X., Yang, D., Wen, M., Koedinger, K., & Rosé, C. P. (2015). Investigating how student's cognitive behavior in MOOC discussion forums affect learning gains. *Conference on Educational Data Mining (EDM).*

Xylogiannopoulos, K., Karampelas, P., & Alhajj, R. (2017). Text mining in unclean, noisy or scrambled datasets for digital forensics analytics. *2017 European Intelligence and Security Informatics Conference (EISIC)*, (pp. 76–83). Athens,.

Yıldırım, F. M., Kaya, A., Öztürk, S. N., & Kılınç, D. (2019). A real-world text classification application for an E-commerce Platform. *2019 Innovations in Intelligent Systems and Applications Conference (ASYU).* Izmir, Turkey.

Yu, F., & Zheng, D. (2017). Education data mining: How to mine interactive Text in MOOCs using natural language process. *The 12th International Conference on Computer Science & Education (ICCSE 2017).*

Zaman, K., & Mamun, K. A. (2017). An evaluation of smartphone apps for preventive healthcare focusing on Cardiovascular Disease. *4th International Conference on Advances in Electrical Engineering (ICAEE),*.

Zhang, M. (2020). E-commerce comment sentiment classification based on deep learning. *IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA).* Chengdu, China.

Abdul Hanan Khan Mohammed received a B. tech degree from Electronics and Communication Engineering, from Jawaharlal Nehru Technological University, Hyderabad, India in 2015. He worked as a Data Analyst at Analytical Instrumentation and Maintenance Systems (AIMS), Abu Dhabi, U.A.E, from 2016 to 2018. He is now pursuing his M.Eng degree in Electrical and Computer Engineering from Ryerson University since 2019. His areas of interest including Data Acquisition, Data Validation, Predictive modelling, Big Data Analytics, RDBMS software (Microsoft SQL Server, Oracle DB, Teradata), Data Visualization, Machine Learning, Text Analytics and Deep Learning Algorithms.

Hrag-Harout Jebamikyous received his Bachelor of Electronics Engineering Technology, Yorkville University, 2018. He is currently a graduate student in the Department of Electrical and Computer Engineering, Ryerson University. His research interests include using machine learning in IoT, Finance, Text Analysis, Deep Learning in Autonomous Vehicles, and Medical Image Analysis.

Dina Nawara is a MASc. student in the Department of Electrical and Computer Engineering at Ryerson University. Prior to that, she has worked in senior telecommunication positions at a number of Telecom leading companies. Her research interests focus on machine learning, and recommendation systems.

Dr. Rasha Kashef received her Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, in 2008. She worked as an Assistant Professor with the School of Computing, AAST Institute, from 2009 to 2011. She worked as a Research Associate at Microsoft Corporation. She worked as a Postdoctoral Fellow with the Department of Applied Mathematics, The University of Waterloo, from 2011 to 2013. She also joined the Department of Management Science, University of Waterloo, from 2013 to 2016. She was hired as an assistant professor with the Management Science Group, Ivey Business School, with a focus on data analytics, from 2016 to 2019. She is currently working as an Assistant Professor with the Department of Electrical, Computer, and Biomedical Engineering, Ryerson University. She is also a Professional Engineer in ON. Her research interest includes machine learning in big data analysis in different applications.