Analysis of Text Mining from Full-text Articles and Abstracts by Postgraduates Students in Selected Nigeria Universities

Mariam Taiwo Ibrahim[1] & Adeyinka Tella[2]

[1] Department of Library and Information Science, University of Ilorin, Nigeria

[2] Department of Library and Information Science, University of Ilorin, Nigeria, and Department of Information Science, University of South Africa, Pretoria, South Africa

Correspondence: Adeyinka Tella, Department of Library and Information Science, University of Ilorin, Nigeria, & Department of Information Science, University of South Africa. E-mail: tella.a@unilorin.edu.ng or tellaa@unisa.ac.za

Received: April 10, 2020 Accepted: June 9, 2020 Online Published: June 10, 2020

doi:10.5430/ijhe.v9n4p169 URL: https://doi.org/10.5430/ijhe.v9n4p169

Abstract

Purpose: This study analysed text mining from full-text articles and abstracts by postgraduate students in selected Nigeria universities.

Design/methodology/approach: The study adopted a survey research design using a questionnaire as the instrument for data collection from 357 postgraduate students drawn using Raosoft sample size calculator. Six research questions were developed and answered.

Finding: The findings demonstrate that postgraduate students mined texts from full texts articles mostly to write a dissertation, for personal academic development and to prepare research seminars. It also revealed that postgraduate students mined texts from abstracts majorly to write dissertations and prepare for research seminars; postgraduate students mined texts using information extraction technique, information retrieval technique, and summarization. The texts are mined mostly form PDF format, followed by Microsoft word format and HTML format (Web pages). Postgraduate students prefer mining texts from full-text articles than from abstracts and the sources postgraduate students mostly mine text is through the World Wide Web, followed by library databases.

Research limitations/implications: The current study only used a questionnaire, a self-reported survey to collect data from the respondents of the study. Including other data collection instruments such as interviews would provide a holistic view of the data mining scenario from both the full-text articles and abstracts among the postgraduate students in Nigerian universities and this would make the generalisation of the study findings easier and more worthwhile.

Originality/value: Research on data mining either from full-text articles or abstracts were predominantly conducted in Advance countries. This study seems to be one of the pioneer studies in this area in Nigeria and Africa as a whole. It is the original idea by the author; and it is assumed that understanding the nature and context-related information in data mining by the postgraduate students is an original idea.

Keywords: data mining, data and information extraction, full text articles, abstracts, postgraduate students, Nigeria

1. Introduction

The burden of the postgraduate programme and the necessity to generate new knowledge, innovations, and research have brought the prerequisite for scholarly and scientific research to a mainstay. This development has provoked the relevance of abstract and full-text articles which are essential in stimulating research and development, most importantly in the academic environment. Not all the textual information in a full-text article can be retrieved compare to abstract. That has necessitated the importance of text mining from full-text articles and abstract to serve the information need of postgraduate students for their research.

Full-text articles are a good source for the latest development in various disciplines. Ordinarily, full -text articles are the reports of academic research that must have undergone a peer-review process to ensure its practicality and provide practical information to practitioners in the field. They are published periodically in journals. Some journals are published annually, bi-annually, and quarterly, bi-monthly, or monthly. Ayeni (2017) observed that postgraduate students use full-text articles for academic purposes as it has become important for scientific research and development. Akpochafo (2009) stressed that universities are charged with the creation of knowledge and it is one of their primary mandates. Conducting research is part of the knowledge creation of which postgraduate students of most involved. Conducting research involves referring to available relevant texts. These texts can be mined from either full-texts articles or abstracts.

According to Gonzalez, Tahsin, Goodale, Greene, and Greene (2016), text mining is a subfield of data mining that involves the extraction of important new information from unstructured (or semi-structured) sources. They expressed further that what distinguishes text mining from natural language processing is that text mining extract information from within those sources (unstructured or semi-structured sources) and aggregate the extracted pieces over the entire collection of source documents to uncover or derive new information. In contrast to Gonzalez et al. (2016) position that text mining is a subset of data mining, Kumar and Bhatia (2013) opined that text mining is similar to data mining except that data mining tools are designed to handle structured data from databases, but text mining can also work with unstructured or semi-structured data sets such as e-mails, text documents, and hypertext mark-up language (HTML) files, etc. They concluded that text mining is a far better technique in the communication of information.

Kumar and Bhatia (2013) defined text mining as the process of analyzing text to extract information that is useful for a specific purpose. They stressed that text mining is the most common way for the formal exchange of information which deals with texts whose function is to facilitate communication of actual information or opinions. De Maio, Fenza, Loia, and Parente, (2017) stressed that the main goal of text mining is to find pertinent information in the full-text and abstract by transforming it into data that can then be analyzed. This is an unavoidable academic activity that is carried out by postgraduate students in the process of getting their class assignments done and most importantly in writing their dissertation or thesis.

Dang and Ahmad (2014) viewed text mining as a multidisciplinary field, concerning retrieval of information, analysis of the text, extraction of information, categorization, clustering, visualization, mining of data, and machine learning. Balamurugan and Pushpa (2015) stressed that text mining is a young interdisciplinary field that is incorporated with data mining, web mining, information retrieval, information extraction, computational linguistics, and natural language processing. Vidhya and Aghila (2010) noted that text mining is also known as knowledge text analysis, text data mining or knowledge-discovery in text. They expressed that text mining and information mining are similar because both procedures “mine” a lot of information, searching for dynamite patterns. They explained further that some of the mining sorts are data, text, web, and business process.

Gaikwad, Chaugule, and Patil (2014) observed that the text mining process involves the collection of text documents from different sources. They expressed that after the collection of these sources, the text mining tool would retrieve a document and pre-process it by checking the format and character sets, after which the document would go through a text analysis phase. This will lead to the creation of new knowledge. Balamurugan and Pushpa (2015) view the process of text mining from document gathering, document pre-processing, text transformation, data mining/pattern selection, and evaluation. Patel and Ghandi (2015) pointed out that the fundamental objective of text mining is to enable users to extract data from text-based assets and manages the operations like information retrieval, classification (supervised, unsupervised and semi-supervised) and summarization. Agrawal and Batra (2013) opined that text mining is a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyze these data objects.

Talib, Hanif, Ayesha, and Fatima (2016) assert that the generic process of text mining performs the following steps. These steps start with the collection of unstructured data from different sources available in different file formats such as plain texts, web pages, pdf files, etc. The second step is the pre-processing and cleansing operations which are carried out to detect and remove anomalies. The third step is the processing and controlling operations which are applied to audit and further clean the data set by automatic processing. The fourth step is the pattern analysis which is implemented by the Management Information System (MIS). The final step is to extract valuable and relevant information for effective and timely decision making and trend analysis.

In the process of text mining from full-text articles and abstracts, there are various techniques used for text mining. Gaikwad, Chaugule, and Patil (2014) explained that to analyze, and generate text; techniques are produced by natural language processing. These techniques which involve information extraction, summarization, categorization, clustering, and information visualization, are used in the text mining process which can be applied to extract knowledge. identified some of the techniques to include natural language processing, machine learning, artificial intelligence, statistical methods, linguistic learning, semantic analysis, predictive modeling, knowledge discovery algorithm, keyword analysis, and classification technique. Liao, Chu, and Hsiao (2012) assert that text mining techniques are continuously applied in industry, academia, web application, Internet, and other fields. This underpins the importance of text mining in the academic or research activities of postgraduate students.

Dang and Ahmad (2014) opined that text mining has become an important research area because large information is stored in different places in an unstructured format. Ramanathan and Meyyapan (2013) found that approximately 80% of the world's data is in unstructured text. They explained further that discovering the patterns and trends in the journals and proceeding from a huge volume of papers is an essential task in the research field. This pool of data cannot be used efficiently except some technique is adopted to extract pertinent and valuable information from the unstructured text which will make a piece of new knowledge. Vidhya and Aghila (2010) argued that text can be sited in emails, chats, SMS, newspaper articles, organization records, journals, and product reviews. This implies that text mining can be carried out on full-text articles or abstracts in the creation of new knowledge.

There are various applications for carrying out text mining activities. Agrawal and Batra (2013) identified text mining applications such as automatic processing of messages and e-mails, analysis of warranty/insurance claims, diagnostic interviews, and analysis of open-ended survey responses. They argued that analysis of close-ended questions cannot be regarded as text mining application because such responses are easily quantified and analyzed, unlike open-ended questions that allow the respondents to answer a question in his/her words. They argued further that such unstructured responses often provide richer and more valued information than close-ended questions and are an important source of insight since they can generate information that was not anticipated. Above all, open-ended survey responses allow for unguided or unrestricted responses and as such will give in-depth information on a phenomenon.

Iarrobino (2017) stated that many researchers use the summary information in abstracts to compile a collection of records for text mining rather than using full-text articles. He observed that there are two reasons for this and the reasons are abstracts are not only easily accessible via various databases, but they also come in a suitable format for the text mining, Extensible Markup Language (XML). The author argued that though using abstracts seems like an easy task, there are major benefits that come from mining full-text articles. Abstracts often do not include all the detailed descriptions of methods, access to essential facts, and all secondary study findings. This indicates that text mining is more possible from full-text articles than abstracts.

2. Statement of the Problem

Text mining from full-text articles and journals is an academic activity that cannot be avoided by postgraduate students in their academic programmes, particularly research. In this study, text mining involves the extraction of text by postgraduate students from full-text articles and abstracts for their various academic study or activity. This helps them to create new knowledge and, in the process, that they can address their academic and information need. Text mining from full-text articles is more reliable than text mining from the abstract (Divoli, Nakov, & Hearst, 2012; Iarrobino, 2017). Nevertheless, it has been observed that scholars "mine" text from abstracts than full-text articles (Iarrobino (2017). Shah, Perez-Iratxeta, Bork, and Andrade (2003) observed that most information extraction from the scientific publications use the abstract of the publication more than the full-text articles because abstracts are available in public databases than full-text articles.

According to Ivwighreghweta and Onoriode (2012), as the use of full-text articles is on the increase, some constraints were identified to have impeded their further use in the literature. Similarly, Dulle (2011) expressed that one of the barriers hindering the use of full-text articles and abstracts is the lack of open access awareness and lack of formal training programs targeted at postgraduate students in the universities. Consequently, postgraduate students most often find themselves spending long hours attempting to acquire crucial and pertinent information from both full-text articles and abstracts than might have been the case if equipped with the required knowledge (Chilimo, 2008; Eger, 2008).

Again, there has been limited study available in Nigeria that has examined text mining from full-text articles and abstracts by university students let alone postgraduate students. Based on the foregoing, it can be observed that there is a large discrepancy in text mining from full-text articles and abstracts by postgraduate students. Also, studies that address the techniques adopted by postgraduate students to mine text from full-text articles and abstracts, and the challenges encounter doing such are currently lacking. Notably, the act of mining text by postgraduate students when carrying out their research is indispensable. Again, it is observed that the results of preference for mining text from full-text articles and abstracts are mixed (Divoli, Nakov, & Hearst, 2012; Iarrobino, 2017). It is against these backdrops, therefore; that this study, seeks to analyze text mining from full-text articles and abstracts by postgraduate students in selected Nigeria universities.

3. Objectives of the Study

The main objective of this study is to analyze text mining from full-text articles and abstracts by postgraduate students in selected Nigeria universities. The specific objectives are to;

i. assess the extent of text mining from full-text articles by postgraduate students in selected Nigeria universities;

ii. examine the extent of text mining from abstracts by postgraduate students in selected Nigeria universities;

iii. identify the techniques used in text mining from abstracts and full-text articles by postgraduate students in selected Nigeria universities;

iv. examine the file formats used in text mining by postgraduate students in selected Nigeria universities;

v. determine the postgraduate students' preference for mining text from full-text or the abstract;

vi. identify the sources where texts are mined by the postgraduate students in selected Nigerian universities.

4. Research Questions

The following research questions will guide this study;

i. What is the extent of text mining from full-text articles by postgraduate students in selected Nigeria universities?

ii. What is the extent of text mining from abstracts by postgraduate students in selected Nigeria universities?

iii. What are the techniques used in text mining by postgraduate students in selected Nigeria universities?

iv. What file formats do postgraduate students use in text mining in selected Nigeria universities?

v. What is the postgraduate students' preference for text mining from full-text articles and abstracts?

vi. What are the sources from which texts are mined by the postgraduate students in selected Nigerian universities?

5. Literature Review

5.1 Text Mining

According to Dang and Ahmad (2014), text mining is a multidisciplinary domain of knowledge that involves the retrieval of information, analysis of the text, drawing out of information, categorization, clustering, visualization, mining of data, and machine learning. Ananiadou, Kumar, and Bhatia (2013) defined text mining as the procedure of examining text to extract information that is of use in a specific purpose. Agrawal and Batra (2013) described text mining as a technique that is concerned with the extraction of information from unstructured and semi-structured data and find a pattern that is new and devoid of what information we have already. According to Stevens (2014), text mining is a multidisciplinary field that is concerned with text analysis, categorization, information extraction, and machine language from natural language texts and databases. For this study, the definition of Stevens (2014) shall be adopted as it is more encompassing and comprehensively touch the grey areas that all scholars used to conceptualize 'text mining'.

5.2 Techniques Used in Text Mining

Sheik (2017) pointed out that text mining involves a methodical and procedural approach in the identification, extraction, and use of text from a text-based document in any format. Irfan *et al.* (2014) observed that the technique involved in text mining multifaceted including information retrieval, topic tracking, text analysis, natural language processing, and information classification based on logical and non-trivial patterns from large data sets. Similarly, Gaikwad, Chaugule, and Patil (2014) also identified the techniques used in the text mining process including information retrieval, information extraction, text summarization, categorization, clustering, and information visualization. It was noted, however, by (Yin, Wang, Qiu, & Zhang, 2007; Xu, Zhang, & Niu, 2008) that most scholars neglect the pre-processing stage which is involved with the simplification of text and is a vital stage in text mining, but it is not an approach in text mining. Text mining techniques are basically used for the identification, extraction, and use of texts. Khan and Xhafa (2009) noted that text mining is an arduous task since it deals with the extraction of texts from unstructured or semi-structured sources. The unstructured form of sources of text mining makes it more a daunting task than data mining. By so doing, text mining requires a procedural approach for it to be accomplished.

5.3 Sources of Text Mining

Patel and Ghandi (2015) averred that the benefits of unstructured and semi-organized data can also be seen in textual information sources such as government electronic repositories, news articles, computerized libraries, electronic mail, World Wide Web, talk rooms, databases, and website archives. According to the University of Queensland (2018), sources of text data include library databases, social media, open sources, web scraping, and language corpora. Penn Libraries (2018) added that there are different sources of text mining at Penn University Libraries. These according to the author include Gale text data, LexisNexis Bulk data, HathiTrust digital library, HathiTrust extracted features, chronicling America, DocSouth collections, and the Chinese text project. Berkeley Library of University of California (2018) noted that there are ten (10) categories of sources of text for text mining. These sources are books, scholarly journals, government documents, linguistic corpora, literature, social media, historical, and archival collections, data repositories, newspapers and magazines, citation, and metadata. Relevant to that, Adamopolous (2014) identified the sources of text mining namely; kaggle, public data sets on AWS, knowledge discovery database cups, Stanford large network dataset collection, New York open data, yahoo! Labs, quandl.com, and University of Edinburg database. Also, Laurence McKinley Gould Library (2017) observed that the sources of text mining include digital collections, government documents, primary source collections, and licensed resources.

Several available related studies were found in the literature that has either addressed text or data mining from abstract or full text-articles. For instance, a comprehensive and quantitative comparison of text-mining in 15 million full-text articles and their corresponding abstracts, which were published between 1823-2016 was conducted by Westergaard, Stærfeldt, Tønsberg, Jensen, and Brunak (2018). These full-text articles and abstracts were downloaded from Medical Literature Analysis and Retrieval System Online (MEDLINE) corpus, PMC corpus, and TDM corpus. The MEDLINE corpus consists of 26,385,631. From there, empty citations were removed from the corrections and duplicate PubMed IDs. For the duplicate PubMed IDs, only the newest entries were kept. This amounted to a total of 16,544,511. Also, in the PMC corpus, a total of 5,807 documents were discarded, which yielded a total of 1,483,120 articles for text mining. Also, TDM corpus comprises 3,335,400 and 11,697,096 full-text articles in portable document format, PDF format, respectively. The study found that full-text articles have the highest number than abstracts for disease-gene associations. Also, the study found that, in terms of comparison, full-texts perform better than abstracts and that access to the full-text articles improved text mining greatly. This is different from the current study because data in the study were harvested from various databases MEDINE, PubMed, IDM corpus, and PMC corpus, whereas, the current study collected data via a self-reported instrument distributed to the postgraduate students to seek for their perception on data mining from the full text and abstract.

Another study on a survey of text mining from the perspective of social media- Facebook and Twitter users; described how studies in social media have used text analytics and text mining techniques for identifying the key themes in the data (Salloum, Al-Emran, Monem, and Shaahan (2017). The survey focused on analyzing text mining studies related to Facebook and Twitter. The study employed archiving service (twimemachine.com) in December 2014, the complete Twitter timelines of 10 academic libraries were considered to collect the dataset for the study. The libraries of 10 highest-ranking universities from the global Shanghai Ranking were chosen for that purpose. The selection requirement of universities includes they must be English-based. The selection was restricted to only one library in case of universities that have more than one library. Noteworthy weaknesses that were found revealed that all the libraries were English-language based and, in the sample, only 10 academic libraries were considered for the analysis. It was also reported that the keywords and phrases' particularization is very helpful in data mining. This study is different from the current one because text mining was data on text mining that was gathered from the social media (Facebook and Twitter) users while the current study gathered its data from the postgraduate students from Nigerian universities.

Text categorization based on a hybrid approach using feature selection and classification techniques was assessed by (Sridharan and Chitra, 2016). The study initiates a new-fangled technique that associates the characteristics text selection and categorization techniques to pace up the text classification process and then about the low-time consumption. Consequently, the study proposes a new method which was basically the combination of feature selection and classification technique that increases and improves the classification accuracy, and feature selection rate. The study demonstrated the effectiveness of the process using systematic assessment and similarity over 13 datasets. The study also found that the projected algorithm did better than the conventional techniques like Best First Search wrapper method and filtered attributed method. While this study used the categorization of data based on a hybrid approach, the current study adopted an empirical analysis to investigate text mining from full texts and abstracts among the postgraduate students.

Irfan *et al.* (2014) carried out a survey on text mining on social media. The study reviews different text mining techniques to discover various textual patterns from the social Web-based applications which provide a platform for the opportunities to establish a relationship among people leading to mutual learning and sharing of valuable knowledge, such as chat, comments, and discussion boards. The study generated data from social networking websites (Facebook, LinkedIn, and MySpace) which are inherently unstructured and fuzzy. Due to the complex nature of the data, the process of analyzing, and extracting information patterns poses a major challenge. It was found that people do not care about the spellings and accurate grammatical construction of a sentence that may lead to all sorts of ambiguities, such as lexical, syntactic, and semantic. It was revealed further that the text analysis approach adopted for the exploration of unstructured text on many social networking sites is classification and clustering techniques. This study seems to be a bit related to the current study in terms of its review of different text mining techniques to discover various textual patterns. However, while it collected data from social networking websites such as Facebook, Linkedln, and MySpace just like others, the current study collected it data from the Nigerian postgraduate students through the self-reported survey.

From the analysis of related studies, it is evident that most of the studies collected data from social media users like Facebook, Twitter, MySpace, Linkedln, and others while the current study collected its data from postgraduate students. Most of the studies focused on text mining from social media websites while the current studies focus on text mining from full-texts and abstracts. None of the studies examined a comparative analysis of text mining from the full text and abstract and most of the studies were conducted outside of Africa which revealed that studies of this nature are either limited or not available at all in the African context. On this note, this study considered it important to make data available in Africa from the population of Nigerian postgraduates' students. Thereby serving as an addition to the literature on data mining research.

6. Methodology

6.1 Research Design

This study adopts a descriptive survey design. A descriptive survey was adopted because it involves describing characteristics of the population of interest. A survey was chosen because it assists the authors to obtain a large amount of data within a short period and to have assistance in administering the questionnaire used for data collection (Abdudaiwi, 2018). Also, a survey was considered appropriate because close-ended questions that were precise and specific in the questionnaire help respondents to supply basic answers while open-ended questions enable them to provide rich data and perhaps introduce other avenues of research to explore. Besides, this study adopted a survey because it is considered a good way to obtain accurate data on data mining from full texts and abstracts, given respondents' free hand to express their opinion. Similarly, survey data are easy to analyse especially with the use of SPSS, survey data are also easy to store and access and because survey data sometimes lead to interesting findings that might not be initially considered by the researcher (Abdudaiwi, 2018; Creswell, 2013, 2014; Creswell & Poth, 2018).

6.2 The Population of the Study

The target population for the study consists of postgraduate students from three universities University of Ilorin, Al-hikmah University, and Kwara State University in Kwara State, Nigeria. The total population of the postgraduate students in these universities as of the 2017/2018 session is 5,222 and the breakdown is presented in table 1.

Table 1. Population of the Study

Universities	Population
University of Ilorin	4795
Al-hikmah University	53
Kwara State University	374
Total	5222

Source: Academic Planning Unit (2018)

6.3 Sample Size and Sampling Technique

This study adopts a proportionate stratified random sampling technique. This sampling technique was adopted to allow for proportionate representation viz-a-viz the sample in each university. In other words, the technique was chosen to enable the equal representation of postgraduate students from these universities. This analysis of how the sample was chosen from each of the three universities presented in table 1. Raosoft sample size calculator was used to arrive at 357 samples for the study.

Table 2. Sample Size in Each University

University	Population	Sample
University of Ilorin	4795	327
Al-hikmah University	53	4
Kwara State University	374	26
Total	5222	357

Source: Author-designed (2019)

6.4 Data Collection Instrument

The instrument used for data collection was a researcher designed questionnaire, titled

“Questionnaire on Text Mining from Full-text Articles and Abstracts” (QTMFTA)”. A Questionnaire has been known as one of the most effective instruments for collecting data in a survey study. The questionnaire was divided into two sections, i.e. Section A and B. Section A collected data on the demographic information of the respondents while Section B features items related to text mining from full-text articles and abstracts. The response was based on a 4-point Likert rating scale ranging from 1= Strongly Agreed to 4= Strongly Disagree. Section B was divided into six parts. Each captured data on each objective of the study.

6.5 Validity and Reliability of the Instrument

To achieve the validity of the questionnaire used for data collection, it was given to colleagues in Library and Information Science to assist in checking for the relevance of the items and possibly of measuring what it was intended to measure. Some items in the instrument were reworded, others removed, and substituted based on the suggestions of colleagues before it was finally administered on the respondents. To achieve the reliability of the questionnaire, a test-retest method was adopted. Copies of the questionnaire were administered on twelve postgraduate students of Caleb University, Imota, Lagos State over a ten days interval. The collected responses were analysed using Pearson Product Moment Correlation to ascertain the coefficient of pre-test and post-test of each segment of the questionnaire. The outcomes yielded the following correlation co-efficient for the sub-scale and the entire scale.

Table 3. Reliability

S/N	Variables	No of Items	Co-efficient
1	The extent of text mining from full-text articles	7	0.79
2	The extent of text mining from abstracts	5	0.68
3	Techniques used in text mining	5	0.69
4	File Formats use in text mining	5	0.70
5	Preference for text mining from full-text articles and abstracts	5	0.69
6	Sources from which texts are mined	5	0.73
	Overall Coefficient	32	0.71

Source: Author’s fieldwork (2019)

6.6 Data Collection Procedure

The researcher requested approval from the appropriate authority in each of the three universities to administer the questionnaire. This allowed the researchers the opportunity to access the respondents. Copies of the questionnaire were administered to the respondents by the researcher and research assistants. They were made to understand that all their responses would be treated with strict confidence. Similarly, they were made to know that they reserve the right to participate or withdraw from the study at will. So also, each of the respondents was accorded their due respect in terms of seeking their informed consent before giving them a copy of the questionnaire. Also, none of the respondents was forced to take part in the filling of the questionnaire but were intimated with the purpose and what the study set to achieve. The respondents willingly and voluntarily participated in the study.

6.7 Method of Data Analysis

This study is descriptive survey research, hence the data collected were analyzed using the descriptive method of analysis and inferential statistics. Data on research questions were analyzed using frequency counts and simple percentages, mean and standard deviation.

7. Results

Based on the sample of the study, a total of 357 copies of the questionnaire was administered based on the sample selected while 329 copies representing 92.2% were returned properly filled and good for data analysis.

Table 4. Demographic Information of the Respondents

Items	Frequency	Percentage (%)
Gender		
Male	144	43.8
Female	185	56.2
Total	329	100.0
Age		
Less than 20 years	11	3.3
20-25 years	156	47.4
26-30 years	121	36.8
31-35 years	16	4.9
36-40 years	10	3.0
41 years and above	15	4.6
Total	329	100.0
Institutions		
Kwara State University	25	7.6
Al-hikmah University	4	1.2
University of Ilorin	300	91.1
Total	329	100.0
Year of Study		
Year 1	157	47.7
Year 2	123	37.4
Year 3	49	14.9
Total	329	100.0

Source: Author’s Fieldwork (2019)

Table 4 shows that 144(43.8%) of the respondents were male students while 185(56.2%) were female students. This means that most of the respondents were female. Also, the table shows that 11(3.3%) of the respondents’ age was less than 20 years, 156(47.4%) were between 20-25 years, 121(36.8%) were between 26-30 years, 16(4/9%) were between 31-35 years, 10(3.0%) were between 36-40 years and 15(4.6%) were 41 years and above. This indicates that most of the respondents were between 20-25 years of age while respondents whose ages were between 36-40 years were the least.

The Table also reveals that 25(7.6%) of the respondents were from Kwara State University (KWASU), 4(1.2%) were from Al-hikmah University and 300(91.1%) were from the University of Ilorin. This simply means that the University of Ilorin had more Master students than KWASU and Al-hikmah University. Moreover, it can be observed that 157(47.7%) of the respondents were in the first year of the postgraduate programme, 123(37.4%) were in the second year while 49(14.9%) were in the third year. This implies that most of the respondents were in the first year of postgraduate programme.

7.1 Analysis of Research Questions

Research Question 1: What is the extent of text mining from full-text articles by postgraduate students?

Table 5. Extent of text mining from full-text articles by postgraduate students

	Responses							
Items	Very High Extent		High Extent		Low Extent		Very Low Extent	
	N	%	N	%	N	%	N	%
Assignments	52	15.8	36	11.0	112	34.0	129	39.2
Personal development	89	27.0	121	36.8	72	21.9	47	14.3
Dissertation	163	49.6	136	41.3	11	3.3	19	5.8
Class exercise	58	17.6	72	21.9	114	34.7	85	25.8
Research seminar	81	24.6	125	38.0	53	16.1	70	21.3

Source: Author’s Fieldwork (2019)

Table 5 shows that 88(26.8%) of the respondents mine text from full-text articles to write their assignments compare to 241(73.2%) mine text from full-text articles to a low extent to write their assignments. The table shows again that, 210(63.8%) of the respondents mine text from full-text articles for personal academic development to a high extent compare to 119(36.2%) who mined text from full-text articles for their personal development to a low extent. 299(90.9%) of the respondents mine text to a high extent from the full-text articles to write their dissertation while 30(9.1%) mine text from full-text articles at a low extent to write their dissertation. Also, 130(29.5%) of the respondents mine text from the full-text articles for their class exercise to a high extent while 199(60.5%) mine text from full-text articles for their class exercise to a low extent. Besides, the table shows that 206(62.6%) of the respondents mine text from the full-text articles to prepare their research seminar to a high extent while 123(37.4%) of the respondents mine text from the full-text articles for their research seminar at a low extent. The results here imply that postgraduate students mined texts from full-text articles mostly to write a dissertation, for personal academic development and to prepare research seminars.

RQ 2: What is the extent of text mining from abstracts by postgraduate students?

Table 6. Extent of text mining from abstracts by postgraduate students

	Responses							
Items	Very High Extent		High Extent		Low Extent		Very Low Extent	
	N	%	N	%	N	%	N	%
Assignments	19	5.8	26	7.9	123	37.4	161	48.9
Personal development	15	4.6	18	5.4	156	47.4	140	42.6
Dissertation	127	38.6	115	35.0	42	12.7	45	13.7
Class exercise	31	9.4	40	12.2	123	37.4	135	41.0
Research seminar	96	29.2	119	36.2	52	15.8	62	18.8

Source: Author’s Fieldwork (2019)

Table 6 shows that 45(13.7%) of the respondents mine text to a high extent from abstracts to write their assignment compare to 284(86.3%) who mined text from abstract at the low extent to write their assignment. The table shows that 33(10.0%) of the respondents mine text from abstracts to a high extent for their personal academic development and 296(90.0%) mine text to a low extent for their personal development. Moreover, the data in the table shows that 242(73.6%) of the respondents mine text from abstracts to a high extent to write their dissertation whereas 87(26.4%)

mine text from abstracts to a low extent to write their dissertation. Also, the table shows 71(21.6%) of the respondents mine text from abstracts to a high extent for their class exercise while 258(78.4%) mined text from abstracts to a low extent for their class exercise. 215(65.4%) of the respondents mined text from abstracts to a high extent to write their research seminar while 114(34.6%) of the respondents mine text from abstracts to a low extent to prepare their research seminar. This indicates that postgraduate students mined texts from abstracts majorly to write a dissertation and to prepare for research seminars.

RQ 3: What are the techniques used by postgraduate students in text mining?

Table 7. Techniques used in text mining

Items	Responses							
	Strongly Agreed		Agreed		Disagreed		Strongly Disagreed	
	N	%	N	%	N	%	N	%
Information retrieval	125	38.0	136	41.3	25	7.6	43	13.1
Information extraction	153	46.5	139	42.2	21	6.4	16	4.9
Natural language processing	49	14.9	52	15.8	107	32.5	121	36.8
Text summarization	56	17.0	62	18.9	98	29.8	113	34.3
Information classification	36	10.9	23	7.0	141	42.9	129	39.2

Source: Author's Fieldwork (2019)

The Table 7 shows that 261(79.3%) of the respondents agreed that postgraduate students mined texts using information retrieval; 292(88.7%) mine text using information extraction while 101(30.7%) mined text using natural language processing technique. Also, 228(69.3%) mine text by summarization. This implies that most of the postgraduate students mined texts using information extraction techniques, information retrieval techniques, and summarization.

RQ 4: What are the file formats postgraduate students mine text from?

Table 8. File formats postgraduate students mine text from

Items	Responses							
	Strongly Agreed		Agreed		Disagreed		Strongly Disagreed	
	N	%	N	%	N	%	N	%
Microsoft word	114	34.7	126	38.3	60	18.2	29	8.8
PDF file	168	51.1	145	44.1	7	2.1	9	2.7
Web pages (html)	27	8.2	31	9.4	124	37.7	147	44.7
E-mail messages	11	3.3	7	2.1	171	52.0	140	42.6
Adobe postscript file	45	13.7	60	18.2	124	37.7	100	30.4

Source: Author's Fieldwork (2019)

Table 8 shows that 240(73.0%) of the respondents agreed that postgraduate students mine text from Microsoft word format; 313(95.2%) mine text from portable document file (PDF); and, 58(17.6%) of the respondents mine text from the webpage. Also, 18(5.4%) mine text from e-mail messages while 105(31.9%) mine text from adobe postscript. This implies that postgraduate students mine text mostly form PDF format, followed by Microsoft word format, while limited percentage mine text from Web pages.

RQ 5: What is the postgraduate students' preference for text mining from full-text articles and abstracts?

Table 9. Use of E-collaboration by Postgraduate Students

S/N	Use of E-collaboration	Freq.	Percentage	Mean x̃	SD
1.	I prefer text mined from a full-text article than abstracts	247	75.1	0.75	13.6
2.	I prefer text mined from abstracts than full-text articles	82	24.9	0.25	4.5
	Total	329	100		

Source: Author's Fieldwork (2019)

Table 9 shows that postgraduate students prefer mining texts from full-text articles than from abstracts. The frequency means and standard deviation of those who prefer mining texts from full-text articles are far higher than the frequency of those who prefer mining texts from abstracts.

RQ 6: What are the sources from which texts are mined by the postgraduate students in selected Nigerian universities?

Table 10. Sources where texts are mined by postgraduate students

Items	**Responses**							
	Strongly Agreed		Agreed		Disagreed		Strongly Disagreed	
	N	%	N	%	N	%	N	%
Library databases	132	40.1	115	35.0	53	16.1	29	8.8
World Wide Web	189	57.4	140	42.6	-	0.0	-	0.0
Social media	23	7.0	41	12.5	140	42.5	125	38.0
Data repositories	59	17.9	61	18.5	96	29.2	113	34.4
Newspapers/Magazines	75	22.8	63	19.1	103	31.3	88	26.7
Citation and metadata	12	3.7	9	2.7	173	52.6	135	41.0
Blogs	89	27.1	57	17.3	76	23.1	107	32.5

Source: Author's Fieldwork (2019)

Table 10 shows that 247(75.1%) of the respondents mine text from library databases; 329(100.0%) of the respondents mine text from World Wide Web and 64(19.5%) mine text from social media. Furthermore, 120(36.4%) mine text from data repositories; 138(41.9%) mine text from newspapers/magazines; 21(6.4%) of the respondents mine text from citation and metadata and146(44.4%) mine text from blogs. This implies that the sources postgraduate students mostly mine text is through the World Wide Web, followed by library databases.

7.2 Discussion of Findings

The study focused on analyze text mining from full-text articles and abstracts by postgraduate students in selected Nigeria universities. The objectives are to assess the extent of text mining from full-text articles by postgraduate students in selected Nigeria universities; examine the extent of text mining from abstracts by postgraduate students in selected Nigeria universities; identify the techniques used in text mining from abstracts and full-text articles by postgraduate students in selected Nigeria universities; examine the file formats used in text mining by postgraduate students in selected Nigeria universities; determine the postgraduate students' preference for mining text from full-text or the abstract, and identify the sources where texts are mined by the postgraduate students in selected Nigerian universities. The findings of the study revealed that postgraduate students mined texts from full-texts articles mostly to write a dissertation, for personal academic development and to prepare research seminars; postgraduate students mined texts from abstracts majorly to write a dissertation and to prepare for research seminars; postgraduate students mined texts using information extraction technique, information retrieval technique, and summarization. The texts are mined mostly form PDF format, followed by Microsoft word format and Web pages. Postgraduate students prefer mining texts from full-text articles than from abstracts and the sources postgraduate students mostly mined text are through the World Wide Web, followed by library databases.

Mining texts from full-text articles mostly to write a dissertation, and for personal academic development was high and postgraduate students highly mined texts from abstracts majorly to write a dissertation and to prepare for research seminars. It is clear from this result that both the abstracts and full-text articles are good for the dissertation and as well for personal development. A researcher may have a robust review of the literature on the theoretical aspect if full details of accessible studies and research are not available. Also, the likelihood of having a robust empirical literature by a researcher in the absence of related abstracts is doubtful. This implies both the full-text articles and abstracts are indispensable when writing a dissertation and when researchers are trying to develop themselves academically.

Postgraduate students mined texts using information extraction techniques, information retrieval techniques, and summarization. As reflected in the related studies, this finding supports the results of some previous studies. For instance, Irfan et al. (2014) identified the technique involved in text mining to include information retrieval, topic tracking, text analysis, natural language processing, and information classification based on logical and non-trivial patterns from large data sets. Similarly, Gaikwad, Chaugule, and Patil (2014) listed the techniques used in the text mining process as information retrieval, information extraction, text summarization, categorization, clustering, and information visualization. All of these agree with the result of this subject in this current study. Also, the emphatic revelation by Sorensen (2009) that text mining involves methodical and procedural approaches in the identification, extraction, and use of text from a text-based document in any format lends credence to the current finding in this study.

Postgraduate students mine text mostly from PDF format, followed by Microsoft word format, while limited percentage mine text from Web pages and mostly prefer mining texts from full-text articles than from abstracts. The result here might be due to the full details information usually provided by full texts than abstract. Just like full articles, abstracts can as well provide insight on a bibliographic compilation which might be useful for researchers to further their research. However, while abstracts are only useful in the empirical area of literature review, full-text articles are useful in both the theoretical and empirical review of the literature. The report by Westergaard, Stærfeldt, Tønsberg, Jensen, and Brunak (2018) on the analysis of full-text articles and abstracts that were downloaded from Medical Literature which revealed that full-texts perform better than abstracts and that access to the full-text articles improved text mining greatly explain the rationale behind the preference for full-text articles in this study than the abstracts.

The sources postgraduate students mostly mine text is through the World Wide Web, followed by library databases. This finding contradicts most of the related findings in the literature. For example, the University of Queensland (2018), Enderle (2018), Laurence McKinley Gould Library (2017), and Adamopolous (2014) all of which identified databases as the most prominent source through which data can be mined. The notion and understanding that most databases now reside on the Web may be the reason why the postgraduate students in this study identified the World Wide Web as the first source through which data can be mined compare to databases.

8. Conclusion

The study based on its findings concludes that postgraduate students mined texts from full-text articles mostly to write a dissertation, for personal academic development and to prepare for research seminars; postgraduate students mined texts from abstracts majorly to write dissertations and to prepare for research seminars; postgraduate students mined texts using information extraction techniques, information retrieval technique, and summarization. The texts are mined mostly form PDF format, followed by Microsoft word format and Web pages. Postgraduate students prefer mining texts from full-text articles than from abstracts and the sources postgraduate students mostly mine text is through the World Wide Web, followed by library databases.

9. Recommendations

The following recommendations were made based on the findings of this study. That most published research papers should be available through Open Access (OA) since the results revealed most postgraduate students prefer to mine text from full-text articles to abstracts for most of their exercise. If open access is not available, access to the full-text articles may not be possible.

It is recommended that journal publishers should endeavour to put their articles or abstracts in PDF Microsoft Word formats, as these formats are most preferable to postgraduate students when they are trying to mine text.

Both full-text articles are abstracts are encouraged to be made open access since this study has reported that both are useful for the postgraduate students in writing dissertation and for personal academic development.

10. Implications and the Contribution of the Study to Knowledge

The findings of this study will be of relevance to data analysts, content analysts, bibliography compilers, journal publishers and database managers as it has determined the file formats postgraduate students prefer to “mine” text from. Also, it has provided insight on text mining from full-text articles and abstracts by postgraduate students.

Making full-text articles and abstracts available through open access will improve the possibility of citations and doing so will improve the h-index, author-level and the productivity, and citation impact of the publication of the respective authors.

Finally, the finding of this study will also be of advantage to future researchers who endeavour to research on text mining from full-text articles and abstracts. This will help to extend the frontiers of knowledge in this research area.

11. Limitations and Suggestions or Future Research

This study has some limitations. For instance, the sample is limited to the postgraduate students selected from three universities in a Nigerian state at the expense of other postgraduate students in other Nigerian universities. Future research should consider extending the scope of the study to cover those zones not involved in this study.

The current study only used a questionnaire, a self-reported survey to collect data from the respondents of the study. Including other data collection instruments such as interviews would provide a holistic view of the data mining scenario from both the full-text articles and abstracts among the postgraduate students in Nigerian universities and this would make the generalisation of the study findings easier and more worthwhile.

Future studies should examine the factors that predict the preference or mining texts from full-text articles than abstracts by postgraduate students.

References

Abduldaiwi, D. (2018). Surveys, advantages, and disadvantages of’ In: Allen Mike (Eds.). *The SAGE Encyclopedia of Communication Research Methods*. SAGE Publications, Inc City: Thousand Oaks.

Adamopolous, P. (2014). *Data sources for data mining and machine learning projects*. Retrieved from http://people.stern.nyu.edu/padamopo/blog/2014-02-20-data-sources.html

Agrawal, R., & Batra, M. (2013). A detailed study on text mining techniques. *International Journal of Soft Computing and Engineering, 2*(6), 118-121.

Akpochafo, W. P. (2009). Revitalizing research in Nigerian universities for national development. *Educational Research and Review, 4*(5), 247-251.

Ayeni, P. O. (2017). Perceptions and use of open access journals by Nigeria postgraduate student. *Journal of Information Science and Theory Practice, 5*(1), 26-46.

Balamurugan, R., & Pushpa, S. (2015). *A review on various text mining techniques and algorithms*. Paper presented at the 2nd International Conference on Recent Innovations in Science, Engineering, and Management, 22 November 2015, JNU Convention Center, Jawaharlal Nehru University, New Delhi.

Berkeley Library of University of California. (2018). *Text mining and computational text analysis: Sources*. Retrieved from http://guides.lib.berkeley.edu/text-mining

Chilimo, W. L. (2008). Training in online search skills at Sokoine University of Agriculture, Tanzania: The use of TEEAL and AGRORA databases. *University of Dar es Salaam Journal, 10*(1&2), 68-80. https://doi.org/10.4314/udslj.v10i1-2.43416

Creswell, J.W. (2013). *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Creswell, J.W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 4th ed. Thousand Oaks. Sage.

Creswell, J.W., & Poth, C.N. (2018). *Qualitative inquiry and research design: choosing among five approaches*. Thousand Oaks. Sage.

Dang, S., & Ahmad, P. H. (2014). Text mining: techniques and its application. *International Journal of Engineering & Technology Innovations, 1*(4), 22-25.

De Maio, C., Fenza, G., Loia, V., &Parente, M. (2017). Text mining basics in bioinformatics. *Journal of LATEX Templates, 2*(8), 1-28.

Divoli, A., Nakov, P., & Hearst, M. A. (2012). Do peers see more in a paper than its authors? *Advanced Bioinformatics,* 750214. https://doi.org/10.1155/2012/750214

Dulle, F. W. (2011). Acceptance and usage of Open Access scholarly communication by postgraduate students at the Sokoine University of Agriculture and the University of Dar es Salaam, Tanzania. *African Journal of Library, Archives and Information Science, 21*(1), 17-28.

Eger, A. (2008). Database statistics applied to investigate the effects of electronic information services on publication of academic research – A comparative study covering Austria, Germany, and Switzerland. *GMS Med Bibl Information, 8*(1), retrieved from http://Www.Egms.De/En/Journals/Mbi/2008-8/Mbi000104.Shtml

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications, 85*(17), 42-45. https://doi.org/10.5120/14937-3507

Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., & Greene, C. S. (2016). Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics, 17*(1), 33-42. https://doi.org/10.1093/bib/bbv087

Iarrobino, M. (2017). *The benefits of text mining full-text instead of abstracts*. Retrieved from http://Benefits%20of%20Text%20Mining%20Full%20Text%20Instead%20of%20Abstracts.html

Irfan, R., King, K. C., Es, D. G., Ewen, S., Khan, S. U., Madani, S. A., Kolodziej, J. O., Wang, L., Chen, D., Rayes, A. R., Tziritas, N., Xu, C. Z., Zomaya, L. Y., Alzaharani, A. S., & Li, H. O. N. (2014). A survey of text mining in social networks. *The Knowledge Engineering Review, 00*(0), 1-24.

Ivwighreghweta, O., & Onoriode, K. O. (2012). Awareness of Open Access scholarly publication among lecturers in university of Benin, Edo State, Nigeria. *Journal of Research in Education and Society, 3*(1), 1-11.

Kaur, A., & Chopra, D. (2016). *Comparison of text mining tools*. Paper presented at the 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Sep. 7-9, 2016, AIIT, Amity University Uttar Pradesh, Noida, India. https://doi.org/10.1109/ICRITO.2016.7784950

Khan, S. U., & Xhafa, F. (2011). *Genetic algorithms for energy-aware scheduling in computational grids*. In Proceedings of 6th IEEE International Conference on P2P, Parallel, Grid, Cloud, and Internet Computing (3PGCIC).

Kumar, L., & Bhatia, P. K. (2013). Text mining: Concepts, process and applications. *Journal of Global Research in Computer Science, 4*(3), 36-39.

Laurence McKinley Gould Library. (2017). *Text mining*. Retrieved from https://gouldguides.carleton.edu/c.php?g=599438&p=4175518

Liao, S. H., Chu, P. H., & Hsiao, P. Y. (2012). Data mining techniques and applications – a decade review from 2000 to 2011. *Expert Systems with Applications, 39*(12), 11303-11311. https://doi.org/10.1016/j.eswa.2012.02.063

Patel, S., & Ghandi, P. (2015). A detailed study on text mining using generic algorithm. *International Journal of Engineering Development and Research, 2*(12), 123-145.

Penn Libraries. (2018). *Text mining at Penn libraries*. https://guides.library.upenn.edu/penntdm

Ramanathan, V., & Meyyapan, T. (2013). *Survey of text mining*. Paper presented at the International Conference on Technology and Business and Management, March 2013, pp. 508-514.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalam, K. (2017). A survey of text mining in social media: Facebook and twitter perspectives. *Advances in Science, Technology and Engineering Systems Journal, 2*(1), 127-133. https://doi.org/10.25046/aj020115

Shah, P. K., Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2003). Information extraction from full-text articles: where are the keywords? *BioMed Central Bioinformatics, 4*(20), 1-9. https://doi.org/10.1186/1471-2105-4-20

Sheikh, T. H. (2017). Text mining and its applications. *International Journal of Allied Practice Research and Review, 4*(11), 1-8.

Sridharan, K., & Chitra, M. (2016). Experimental Investigation for Text Categorization Based on Hybrid Approach Using Feature Selection and Classification Techniques. *Asian Journal of Information Technology, 15*, 2355-2366.

Stevens, D. (2014). *Predicting real estate price using text mining: Automated real estate description analysis*. (Master thesis, Department of Communication and Information Sciences, Tilburg Centre for Cognition and Communication).

Talib, R., Hanif, H. K., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications, and issues. *International Journal of Advanced Computer Science and Applications, 7*(11), 414-418. https://doi.org/10.14569/IJACSA.2016.071153

The University of Queensland Library. (2018). *Text mining and text analysis*. Retrieved from https://guides.library.uq.edu.au/research-techniques/text-mining-analysis/sources-of-text-data

Vidhya, K. A., & Aghila, G. (2010). Text mining process, techniques, and tools: An overview. *International Journal of Information Technology and Knowledge Management, 2*(2), 613-622.

Westergaard, D., Stærfeldt, H., Tønsberg, C., Jensen, L. J., & Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLOS: Computational Biology, 14*(2), 1-16. https://doi.org/10.1371/journal.pcbi.1005962

Xu, X., Zhang, F., & Niu, Z. (2008). *An ontology-based query system for digital libraries*. In Proceedings of IEEE, Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Wuhan, 222-226. https://doi.org/10.1109/PACIIA.2008.360

Yin, S., Wang, G., Qiu, Y., & Zhang, W. (2007). *Research and implement of classification algorithm on web text mining*. In Proceedings of 3rd International Conference on Semantics, Knowledge and Grid, China, 446-449. https://doi.org/10.1109/SKG.2007.105. https://doi.org/10.1109/SKG.2007.249