doi: 10.15678/ZP.2021.55.1.02

Zarządzanie Publiczne
Public Governance
No. 1(55)/2021
ISSN 1898-3529

Dwijendra Nath Dwivedi, Abhishek Anand

The Text Mining of Public Policy Documents in Response to COVID-19: A Comparison of the United Arab Emirates and the Kingdom of Saudi Arabia

Abstract

Objective: The objective of the paper is to analyse publicly available government policy documents of the United Arab Emirates (UAE) and the Kingdom of Saudi Arabia (KSA) in order to identify key topics and themes for these two countries in relation to the COVID-19 response.

Research Design & Methods: In view of the availability of large volumes of documents as well as advancement in computing system, text mining has emerged as a significant tool to analyse large volumes of unstructured data. For this paper, we have applied latent semantic analysis and Singular Value Decomposition (SVD) for text clustering.

Findings: The results of the analysis of terms indicate similarities of key themes around health and pandemic for the UAE and the KSA. However, the results of text clustering indicate that focus of the UAE' documents in on 'Digital'-related terms, whereas for the KSA, it is around 'International Travel'-related terms. Further analysis of topic modelling demonstrates that topics such as 'Vaccine Trial', 'Economic Recovery', 'Health Ministry', and 'Digital Platforms' are common across both the UAE and the KSA.

Contribution / Value Added: The study contributes to text-mining literature by providing a framework for analyzing public policy documents at the country level. This can help to understand the key themes in policies of the governments and can potentially aid the identification of the success and failure of various policy measures in certain cases by means of comparing the outcomes.

Implications / Recommendations: The results of this study clearly showed that text clustering of unstructured data such as policy documents could be very useful for understanding the themes and orientation topics of the policies.

Keywords: text mining, COVID-19, public policy, information extraction, topic modelling, text clustering

Article classification: research paper

JEL classification: D78, E61, I18, L38

Dwijendra Nath Dwivedi – EMEA AI and IOT Leader at SAS Institute; Dubai, UAE; e-mail: dwivedy@gmail.com; ORCID: 0000-0001-7662-415X. **Abhishek Anand** – Leader at the Credit Risk Team, HSBC; Kraków, Poland; e-mail: abhishek.igidr@gmail.com; ORCID: 0000-0002-9880-225X.

Introduction

The 2019 Coronavirus disease outbreak (COVID-19) was one of the most significant global challenges of this century for humankind. The pandemic has significant effects on public health, economics, politics, and society (Cheng et al., 2020). Most governments have responded to the COVID-19 outbreak by adopting proactive lockdown measures and conducting robust education campaigns. The containment measures have resulted in a widespread economic collapse with significant impacts on output and employment, and a serious impact on all industries as a result of the sharp drop in consumption. The most developed countries have seen their positive economic growth dip into the red with marked rises in unemployment and an increase in social inequalities (Carracedo et al., 2020). In such a scenario, governments should play a central role in managing the crisis and recovering the economy.

In response to the COVID-19 pandemic, the governments published a lot of information in public through government websites as well as other sources. However, the huge volume of documentation from multiple sources of information can be difficult to track. Hence, it is vital to identify how an analysis can be done quickly and efficiently on the available literature to understand the key themes for COVID-19 and government policy priorities to counter the pandemic. A large number of research papers and case studies have already appeared in major international journals and publications, where researchers have used text mining techniques to identify the process and framework for rapidly performing reviews of large volume of coronavirus studies and publications, and to classify the key research themes for COVID-19 (Cheng et al., 2020).

In recent years, the responsibilities of public health policy have grown beyond reduction and control of infectious diseases, and public health policies have been implemented to tackle other emerging threats like tsunamis and SARS (Shi et al., 2009). In this context of growing importance of public health policy, there is a need to understand the public health policy regarding COVID-19 and various other policy measures taken by the government to counter the pandemic. However, due to large volume of texts and documents available on COVID-19, it is difficult to apply traditional methods of data analysis to understand the key themes of public policy. Hence, we use text mining approach, which helped us analyse unstructured data and focus on two neighbouring countries, namely the UAE and the KSA. The reason behind selecting these two countries is that they share strong political and cultural ties as well as they had a similar peak for COVID-19 cases in 2020. It is therefore interesting to examine how both countries have approached the pandemic through public policy. The purpose of this paper is to analyse public policy texts and documents available through government sources in the UAE and the KSA in response to the COVID-19 pandemic, and then understand the key themes using the above-mentioned text mining approach. The aim of the research is to respond to the main question:

- What are the important themes that emerge from the United Arab Emirates' and the KSA's public policy documents in the context of the COVID-19 pandemic?

Taking into account the above primary research question, the secondary research question for this study is:

- What are the similarities and differences between the public policy approaches of the United Arab Emirates and the KSA with regard to addressing COVID-19, based on the results of textual exploration?

The data for the analysis was extracted from the public websites of the two countries. Public policies published on the websites of the National Department of Disaster Preparedness as well as other websites were considered suitable for text analysis. For this paper, we consider 108 documents for the UAE and 76 documents for the KSA, published between March 2020 and November 2020. We have used the text mining technique, which has gained significant prominence in recent times for analysis of textual documents due to its ability to handle large amounts of data – also unstructured data – which traditional data mining techniques are not capable of.

Text mining has become a significant research field especially after the arrival of big data tools, which can deal with unstructured data. Text Mining is the process of finding new information that had previously been unknown, automatically extracting information from different written or published sources (Gupta & Lehal, 2009). While text mining is similar to data mining, there is a significant difference. Namely, data mining tools are designed to handle structured data from databases, whereas text mining can work with unstructured or semi-structured datasets, such as emails, PDF files, word files, and HTML files. Topic modelling is a text search tool widely used to detect semantic structures hidden within a body of text. It analyses an enormous collection of documents and groups the words into a group of words in order to identify subjects based on the similarity approach.

Two of the most popular methods for topic modelling include Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). LSA analyses the association between a set of documents and their terms using a grouping approach assuming that similar terms and documents will be grouped together. LDA, on the other hand, assumes that each document comprises of a small number of topics and that each word can be attributed to one of the topics contained within the document. The LDA method is an improved Probabilistic Latent Semantic Analysis (PLSA), pioneered by Hofmann (2001), which assumes that each subject is generated by a word probability distribution. For our research, we have used the LSA method, which typically operates on the term-by-document matrix and uses a well-known mathematical matrix decomposition technique called Singular Value Decomposition (SVD) to break down the original data into linearly independent components (Chakraborty et al., 2013).

The results of our analysis indicate the similarity of the approaches of the UAE' and the KSA's governments on the basis of published papers and selected themes. First, we analysed the key terms for the UAE and the KSA by word frequency distribution and word cloud, and the results indicated few common terms for the two countries. Second, the paper also analysed clusters produced by text mining for the United Arab Emirates and the KSA, and some of the clusters highlighted common themes in both sets of corpus. Third, we used topic extraction to identify key topics for the UAE and the KSA from the documents, and we found common topics between the two sets of documents. Overall, it can be concluded that, based on the textual analysis of the literature published by the UAE' and the KSA's governments, there are similarities in key terms and themes.

The remainder of this document is structured along the following lines. Section 2 presents the literature review, highlighting the evolution of topic modelling and also showcasing some of the key research outcomes and different methods applied in this area. Section 3 presents the database and analytical methodology, and Section 4 demonstrates the main results of the research. Finally, Section 5 outlines the most significant findings.

Literature review

The problem related to text mining has been addressed in the past. However, in recent years there has been a lot more focus on text mining due to advancement in computing and the development of capabilities to handle unstructured data. In order to manage the explosion of electronic document archives, new techniques or tools are required to deal with organising, searching, indexing, and reviewing large collections of data in a time efficient manner (Alghamdi & Alfalqi, 2015). As a generalisation, there are two broad approaches to process text, namely Natural Language Processing (NLP) and statistics-based programmes such as topic modelling (Hofmann, 2001). Unlike NLP methods that identify parts of speech and grammatical structure, statistical models such as topic models rely heavily on the „bag of words"

(BoW) hypothesis. In BoW models, the collection of textual documents is quantified in a document-term matrix (DTM), which counts the occurrence of each word (columns) for each document (rows). In the case of most topic models – such as LDA – the DTM is one of two model inputs, along with a number of topics (Wesslen, 2018).

For most document collections, DTM is often very large and sparse. This makes it difficult to use this matrix directly in clustering or in any other algorithms. Therefore, the idea is to reduce the dimensionality of the data while retaining the meaningful information. This has been one of the early motivations for topic modelling. Deerwester et al. (1990) presented one of the first topic models using latent semantic analysis (LSA) and Singular value decomposition (SVD), in which a large DTM is decomposed into a set of about 100 orthogonal factors, from which the original matrix can be approximated by linear combination. They assumed the presence of some underlying latent semantic structure and used statistical techniques to estimate this latent structure.

Hofmann (2001) used unsupervised learning technique called Probabilistic Latent Semantic Analysis (PLSA) by means of adding a probabilistic component to the LSA model and assuming that each topic is generated by a word probability distribution. The advantage of PLSA is that standard statistical methods can be applied for model fitting, model selection, and complexity control. For example, one can assess the quality of the PLSA model by measuring its predictive performance, e.g. with the help of cross-validation. Blei et al. (2003) extended the PLSA model of Hoffman to build the LDA model, which includes a second probability component for the document level. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics. Each topic is, in turn, modelled as an infinite mixture over an underlying set of topic probabilities (Madhoushi et al., 2015). Using these probability distributions, any word can be rank-ordered by each topic in order to determine what the most common word is used when referring to each topic.

The significance of topic modelling lies in detecting the patterns of word-use and then connecting documents which have similar patterns. Essentially, the documents contain a mixture of topics and each topic is a probability distribution over words (Alghamdi & Alfalqi, 2015). Asmossen and Moller (2019) presented a framework to leverage the thematic modelling technique in order to conduct an exploratory literature review of a broad collection of papers. The framework proposed by them enables a large volume of documents to be reviewed in a transparent, efficient, and reproducible way using the LDA method. In general, there are two approaches to document-processing: supervised learning and unsupervised learning. Supervised learning involves the manual coding of a collection of documents prior to conducting an analysis, which requires considerable time to achieve the outcome. On the other hand, unsupervised learning methods, such as topic modelling, do not have the pre-requisite to manually code the documents, which makes it possible to save a lot of time for an exploratory review of a large collection of papers.

Gotipati et al. (2018) used topic modelling and data visualisation to analyse student feedback from seven undergraduate courses taught at the Singapore Management University. They assessed rules-based methods and statistical classifiers for extracting the topics. Mei et al. (2007) used text mining to analyse sentiments of the users on Weblogs, and proposed a probabilistic mixture model called Topic Sentiment Mixture (TSM), where words are sampled by a mixture model of background language, topic language, and two sentiment language models. They present a mechanism for extracting sub-topics, assigning each sub-topic with a positive or negative feeling and evaluating how opinions about a topic change over time. Al-Obeidat et al. (2018) proposed a sandbox for extracting subjects by means of analysing feelings for extracting subjects and their associated feelings in a database. They used LDA for the extraction of subjects as well as the „bag-of-words" feeling analysis algorithm, where polarity is determined according to the frequency of occurrence of positive/ negative words in a document.

Benedetto and Tedeschi (2016) highlight common approaches to analysing feelings in social media streams and related problems with the help of cloud computing. Big data is divided into four features, namely four V's of big data – volume, velocity, variety, and veracity. Volume is the largest amount of data that needs to be stored and handled. Velocity is the frequency of the incoming data. Variety describes different types of data, while veracity refers to the reliability and accuracy of available data.

Text mining in general, and topic modelling in particular, has gained even more prominence in recent post-COVID-19 times, as there has been a plethora of research papers, documents, government publications, and social media information related to the topic. Cheng et al. (2020) presented an overview of the coronavirus literature using a text mining technique, and identified the research themes as well as representative literature for each theme. They used the LDA approach to analyse 7,909 scholarly articles from 1,461 journals. Carracedo et al. (2020) identify current research themes related to COVID-19 as well as their impact on the business community using the Text Mining methodology. Goel et al. (2021) aim to analyse the COVID-19 situation in India and explain possible impacts of policy changes and technological changes. They analysed data, including publications from government sources and media reports, with a focus on policy and technological responses. Sharma et al. (2020) reviewed Twitter data from a hundred NASDAQ companies and provided important insights on key post-COVID-19 supply chain issues that the companies are facing. Using text mining tools, they extracted those themes from Twitter data that concern problems encountered by companies as well as strategies adopted by them. They observe that businesses face many challenges in the post-COVID era due to the mismatch between demand and supply, technological gaps, and the lack of a resilient supply chain.

This paper contributes to the literature on text mining by providing an approach to analysing huge volumes of material in the field of public policy at the national level. At the same time, it also enhances public-policy-related literature by applying the text mining approach in order to discover themes across public-policy documents, which is extremely useful when dealing with huge volume of documents.

Data and methodology

Data for this paper has been taken from public government websites of both countries. Public policy as published on national disaster response ministry and other news sites has been taken as such for text analytics. Various government departments had been publishing the action documents as a part of the COVID-19 response on their websites. We have extracted all such text documents for concerned ministries in the UAE and the KSA. The time frame for extracting the policy documents was March 2020 to November 2020. The list of all the websites which have been the prime source for the documents is provided in the Appendix.

Topic extraction discovers keywords in documents that capture the recurring theme of the text and as such is widely used to analyse large sets of documents for identifying the most common topics in an easy and efficient way. In terms of methodology, we have applied latent semantic analysis and Singular Value Decomposition for text clustering. Clustering divides observations in a dataset into different clusters or groups so that the observations within a group are similar and the observations between the groups are dissimilar. In the text-mining context, clustering divides the collection of documents into various groups based on the presence of similar themes. The algorithm generates clusters based on the relative positioning of documents in the vector space. LSA (also known as Latent Semantic Indexing, or LSI) is a dimensionality reduction technique that typically operates on the term-by-document matrix by way of using a mathematical matrix decomposition technique called Singular Value Decomposition (SVD), which breaks down the original data into linearly independent components. A term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection, while columns correspond to terms.

Mathematically, a full SVD does the following: Consider that A (mXn) is the term-by- document matrix with m>n (more terms than documents) where the entries in the matrix are real numbers (such as the presence or absence of a term, entropy weight, etc.) SVD computes matrices U, S, and V so that the original matrix can be re-created using the formula A = USVT. In this formula, the following is true:

- U is the matrix of the orthogonal eigenvectors of the square symmetric matrix AA^T;
- S is the diagonal matrix of the square roots of the eigenvalues of the square symmetric matrix AA^T;
- V is the matrix of the orthogonal eigenvectors of the square symmetric matrix A^TA.

The fundamental idea of applying classical data-mining techniques to topic modelling relies on transforming text data (unstructured) to numbers (structured). This numerical representation of the text takes the form of a spreadsheet-like structure called a term-by-document matrix. In this matrix, dimensions are determined by the number of documents and the number of terms in the corpus (Chakraborty et al., 2013).

Results and discussion

In this section, we compare the topic modelling results for the UAE and the KSA.

Terms by document report

Both Table 1 and Table 2 show the frequencies of the most relevant terms for the UAE and the KSA. Here, we present top ten relevant terms in order to focus on the key themes which appeared in the documents from the UAE' and the KSA' sources.

From Table 1 and Table 2, one can discover that some of the top terms for both the UAE and the KSA are common, including 'COVID-19', 'health', and 'August', which indicate common themes across documents from both countries. However, in terms of differences, the UAE has the 'announce' verb occurring frequently, while it does not feature

Table 1. Top ten terms for the UAE

Term	Frequency	# Documents
COVID-19	51	36
+ announce	37	36
June	23	18
health	20	12
government	19	16
+ test	19	13
August	18	16
new	18	15
+ pandemic	16	16
+ resident	16	10

Source: Government policy documents listed in the Appendix.

Table 2. Top ten terms for the KSA

Term	Frequency	# Documents
COVID-19	34	31
coronavirus	23	23
health	23	18
ministry	19	17
pandemic	14	14
August	14	13
July	13	13
September	13	12
international	13	9
+ vaccine	13	7

Source: Government policy documents listed in the Appendix.

in top ten terms for the KSA. Similarly, the KSA has ‘vaccine’ as a top ten term in their documents, while it does not feature in top ten terms for the UAE.

Word cloud

A word cloud (also known as a tag cloud or word art) is a simple visualisation of data in which words are shown in varying sizes, depending on how often they appear.

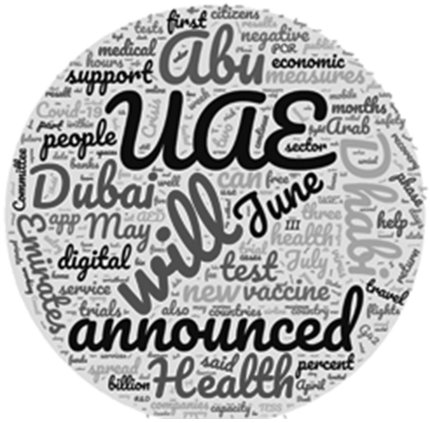

Figure 1. Word cloud of the UAE’ documents

Source: Government policy documents listed in the Appendix.

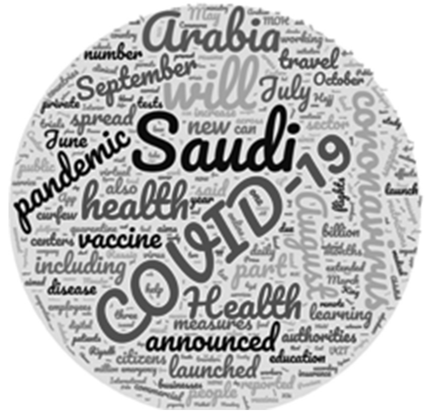

Figure 2. Word cloud of the KSA’s documents

Source: Government policy documents listed in the Appendix.

For the UAE, the largest words are ‘UAE’, ‘June’, ‘Abu Dhabi’, ‘Dubai’, ‘announced’ and ‘Health’. The documents considered for the UAE show that there are a lot of announcements made by the UAE’ government, focusing on Dubai and Abu Dhabi, two main cities of the UAE. For the KSA, the largest words are ‘COVID-19’ and ‘Saudi Arabia’, which is expected. The other larger words – ‘ministry’, ‘health’, ‘pandemic’, ‘July’, ‘August’, and ‘announced’ – demonstrate similar focus for the KSA to that seen for the UAE. One noticeable difference is that for the UAE, the largest month word is ‘June’, whereas for the KSA it is ‘July’ and ‘August’, which could imply that the UAE took many measures in June, whereas the KSA did the same in July and August.

Top clusters and their descriptions

Using the default settings (low SVD resolution and maximum cluster) for the SAS Text Miner, we get 9 clusters for the UAE and 11 clusters for the KSA. By default, the SAS Text Miner uses 15 descriptive terms that best describe each cluster. Table 3 and Table 4 both describe the clusters for the UAE and the KSA.

As can be seen from Table 3, the largest cluster for the UAE is cluster ID 6 containing 19% of total documents, which is related primarily to digital platforms. For the KSA, the largest cluster is cluster ID 6 containing 14% of total documents, and it is related primarily to international travel. If SVD resolution is set too high, then only 5 clusters are generated, as shown in Table 5 and Table 6 for the UAE and the KSA respectively. We have also given a name to each cluster based on the key terms in the cluster; it represents the key theme for the terms in that cluster.

For the UAE, the largest cluster is related to ‘Digital’; it represents 31% of total of the UAE’ documents, whereas for the KSA the largest cluster is ‘International Travel’, which represents 28% of total of the KSA’s documents. Interestingly, the theme of the largest cluster in both cases (low SVD resolution and high resolution) is same for

Table 3. Descriptive terms for clusters for the UAE

Cluster ID	Descriptive Terms	Frequency	Percent-age
1	enter +Emirate Abu Dhabi +development local support +hour +effort spread three +bank +extend +leave +solution	14	13
2	clinical +'clinical trial' +inactivate cnbg iii +trial +vaccine Dhabi-based g42 phase world artificial group intelligence 'artificial intelligence'	7	6
3	+reopen +mall +sector +shop +allow economic capacity department +business June Dubai +begin +permit five people	6	6
4	adgm +fee percent +space +company +partnership +incentive +customer retail +include continue financial +aim +introduce april	9	8
5	period +month +fine three +leave +country August June 'coronavirus pandemic' +conduct +effect +extend +facilitate +permit +flight	9	8
6	affairs social tra +service authority +platform +bank public +resident ministry +guideline management website UAE +launch	20	19
7	recovery package economy +entity billion +business +include support economic government +support +day April week 'coronavirus pandemic'	14	13
8	+passenger travel Arab Emirates United number +citizen +restriction +test +country October million pcr coronavirus precautionary	14	13
9	+result negative +test +require +technology +'COVID-19 test' +disease +identify +solution present medical Dhabi +app. help Abu	15	14

Source: Text clustering on the government documents listed in the Appendix.

Table 4. Descriptive terms for clusters for the KSA

Cluster ID	Descriptive Terms	Frequency	Percent-age
1	+distance authority +area learning +launch +aim +contract king riyadh June people +centre spread September ministry	4	5
2	+platform +school education learning +service remote virtual year +area +distance +prepare +company help March virus	5	7
3	+hospital +citizen +patient app medical +contract +register +resident electronic travel health international +return moh +curfew	9	12
4	'coronavirus disease' disease +'private sector' private +business +sector +extend government coronavirus august +conduct +continue +flight +payment three	6	8
5	+employee +increase remote public +authority +sector +month +country August +area +development +restriction +return help March	6	8
6	number +curfew October travel +development +flight +step moh +report June +start international +authority +citizen +increase	11	14
7	+payment +programme help +extend three +business +month +continue +initiative +recover +service September +start +work +effect	4	5
8	hajj +pilgrim +measure +restriction 'coronavirus pandemic' +doctor pandemic spread coronavirus year +report 'COVID-19 pandemic' +continue +return +step	10	13
9	+vaccine +trial vaccine +develop +work +country +prepare king disease +company +test COVID-19 Arabia people Saudi	7	9
10	+recover July +conduct +initiative people 'coronavirus pandemic' +test +effect Arabia ministry government +month June +flight +patient	9	12
11	digital launch 'COVID-19 pandemic' pandemic virtual +report +aim +initiative +prepare +resident +step moh riyadh +launch +company	5	7

Source: Text clustering on the government documents listed in the Appendix.

Table 5. Descriptive terms for 5 clusters for the UAE

Cluster Name	Cluster Description	Frequency	Percentage
Digital	launched tests operating prevention +department +development +identify health October Abu +platform +technology website national Dhabi	33	31
International Travel	allowed +enter +result entities flights July negative Dubai Abu Dhabi coronavirus +test affairs citizens guidelines	23	21
Vaccine Trial	cnbg clinical percent +fee group +vaccine Dhabi-based g42 iii inactivated companies phase world artificial intelligence	22	20
Economic Recovery	months fines period three Arab introduced affected united package +leave aimed banks incentives +pandemic +include	17	16
Precautionary Guidelines	measures precautionary ,precautionary measures' sectors +'precautionary measure' +country +measure +restriction +support countries COVID-19 +pandemic businesses help restrictions	13	12

Source: Description of text clusters after topic modelling for the UAE.

Table 6. Descriptive terms for 5 clusters for the KSA

Cluster Name	Cluster Description	Frequency	Percentage
International Travel	pandemic electronic ,coronavirus disease' disease extended flights government year +continue +effect +programme international travel coronavirus citizens	21	28
Vaccine Trial + Economic Recovery	private +'private sector' +sector +trial countries trials vaccine developing authority businesses +vaccine developed measures announced disease	17	22
Digital + Education	areas +distance distancing +increase learning launched dr. hajj public virus pandemic September 'COVID-19 pandemic' aimed education	17	22
Digital + Test	people app centers +center conducting recovered tests virus +company +platform +register +test +vaccine aimed health	11	14
Government Services	increased services +increase +step contracting October number launch moh July +month +platform conducted dr. tests	10	13

Source: Description of text clusters after topic modelling for the KSA.

both the UAE and the KSA. For the UAE, the theme is 'Digital', whereas for the KSA, the theme is 'International Travel'. The top 3 clusters for both the UAE and the KSA have similar themes, namely: 'Digital', 'International Travel', and 'Vaccine Trial', although rankings of these clusters are different for the two countries. The contribution of the top 3 clusters for both the UAE and the KSA is 72%, which implies that the major part of policy measures considered by both countries has been revolving around these three themes.

Topic extraction

A topic is a collection of terms that define a theme or an idea. Every document in the corpus can be given a score that represents the strength of association for a topic. A document can contain zero, one, or many topics. The objective of creating a list of topics is to establish combinations of words that are of interest in the analysis (Chakraborty et al., 2013).

We have used the Text Topic Node in the SAS Text Miner in order to discover topics from a text. The node first calculates term topic weight as well as document topic weight. For example, if there are 10 topics extracted, there will be 10 term topic weights calculated for a single term. Similarly, there will be 10 document topic weights calculated for a single document. Term topic weights and document topic weights are then used to calculate cutoff scores for each multi-term topic. Term cutoff is the threshold that determines whether a term belongs to a topic, while document

Table 7. Topics table from the Text Topic Node results for the UAE

Category	Topic Id	Document Cutoff	Term Cutoff	Topic	# Terms	# Docs
Multiple	1	0.205	0.109	+test+resultnegativepresent+require	15	10
Multiple	2	0.207	0.108	+trial+vaccineclinicaliii+inactivate	14	9
Multiple	3	0.16	0.111	+launch+platformwebsite+incentiveapril	24	14
Multiple	4	0.173	0.111	researchvirus+identifymedicalcovid-19	18	10
Multiple	5	0.156	0.11	+residentpermitentryauthority+citizen	15	9
Multiple	6	0.19	0.109	+reopencapacity+sectordubai+allow	16	10
Multiple	7	0.197	0.103	adgm+feepercent+company+space	7	9
Multiple	8	0.21	0.108	healthministry+servicehealth+disease	15	12
Multiple	9	0.181	0.106	+monththreeperiod+leave+fine	10	8
Multiple	10	0.164	0.109	arabunitedemirates+reporteconomic	13	9
Multiple	11	0.193	0.11	billioneconomysupportdubaipackage	19	13
Multiple	12	0.174	0.111	+app.+companydigitalfreeartificialintelligence	19	12
Multiple	13	0.177	0.11	people+enter+emirate+hourjune	18	9
Multiple	14	0.159	0.11	managementcrisisoctober+spacecommittee	17	7
Multiple	15	0.16	0.11	+customer+bankfree+service+platform	16	8
Multiple	16	0.167	0.11	precautionary+measure+precautionarymeasureaffairs	17	10
Multiple	17	0.173	0.109	+day+restriction+measurecontinueweek	14	9
Multiple	18	0.145	0.111	dhabiabudepartmenteconomic+shop	15	11
Multiple	19	0.17	0.11	economicseveralrecovery+support+sector	14	9
Multiple	20	0.168	0.11	+passenger+flightdubaijune+test	14	10

Source: Text topic extracted from the documents listed in the Appendix for the UAE.

Table 8. Topics table from the Text Topic Node results for the KSA

Category	Topic Id	Document Cutoff	Term Cutoff	Topic	# Terms	# Docs
Multiple	1	0.229	0.139	+vaccine +trial +develop vaccine +country	8	8
Multiple	2	0.216	0.139	+sector private +private sector ministry +announce	9	5
Multiple	3	0.208	0.141	+launch+distance+aimauthoritylearning	9	6
Multiple	4	0.221	0.139	+flightinternationaltravel+restrictionseptember	7	6
Multiple	5	0.197	0.14	+curfewjune+authority+returnnumber	7	8
Multiple	6	0.193	0.141	numberoctobermohministrylaunch	10	8
Multiple	7	0.207	0.14	hajj+pilgrimcoronaviruspandemicpandemicpublic	10	8
Multiple	8	0.196	0.139	+test+centerjuly+conductsaudi	7	7
Multiple	9	0.205	0.14	coronavirusdiseaseaugustdiseasespread+citizen	7	7
Multiple	10	0.204	0.141	+hospital+patientmedical+citizenapp	12	9
Multiple	11	0.206	0.14	education+schoollearning+platform+service	11	6
Multiple	12	0.201	0.141	+servicehealthministryhealthjune	11	6
Multiple	13	0.18	0.141	kingdom+measuredigitalgovernment+initiative	12	9
Multiple	14	0.207	0.141	three+programme+payment+extend+month	12	6
Multiple	15	0.175	0.141	+employee+increase+increaseremote+restriction	12	8

Source: Text topic extracted from the documents listed in the Appendix for the KSA.

cutoff is the threshold that determines whether a document belongs to a topic. Table 7 and Table 8 both show the results of topics extraction for the UAE and the KSA respectively. Only the top five weighted terms for each topic are shown in the Topic columns.

As presented in Table 7 and Table 8, there are 20 terms extracted for the UAE and 15 topics extracted for the KSA. Most of the topics and terms extracted for the UAE and the KSA have similarities between them. Topic Id 1 for the KSA refers to vaccine and trial, which is similar to topic Id 2 from Table 7 for the UAE, indicating a broad focus of both governments on clinical trials for a COVID-19 vaccine. The KSA's Health Ministry announced its collaboration with the CanSino Biologics, a Chinese vaccine company, after they had successfully conducted phase I/II trials within China; they also conducted a vaccine trial on a sample population in the KSA (Raja et al., 2020). The KSA also collaborated with Russia on the Gam-COVID-Vac-Lyo, a COVID-19 vaccine currently being developed in Russia. On the other hand, the UAE approved a vaccine developed by Chinese state-owned Sinopharm in December 2020.

Topic Id 2 from the KSA and topic Id 19 from the UAE indicate government measures for reviving economy by supporting private sector. The UAE' initiatives to counter the spread and impacts of the pandemic are coordinated through the UAE' National Emergency Crisis and Disaster Management Authority (NCEMA). These initiatives are multi-sectoral and involve increasing health sector capacity, reviving the private sector through economic incentives, and supporting residents through various programmes and measures. The UAE implemented the 'Targeted Economic Support Scheme' in March 2020 in order to provide support to private companies in the UAE through a series of reliefs. Saudi Arabia was among the first countries to implement early precautionary measures to prevent COVID-19 or to mitigate its impact when it arrives. A national committee was formed in early 2020, consisting of the government ministers for Health, Education, FDA, Interior, and many others in order to fight against the pandemic. The KSA's government also announced a set of support packages in 2020, targeting the private sector and totalling almost $61 billion. In addition, the SAMA (Saudi Central Bank) has been in a continuous dialogue with local commercial banks to support those sectors that are highly impacted by the ongoing pandemic (KPMG Report, November 2020).

Topic Id 8 from the UAE and topic Id 12 from the KSA show the involvement of the Health Ministries of both countries in dealing with the pandemic. Although the UAE has a robust public health system, the pandemic has highlighted the need for improvement in the existing system. The UAE has only 1.3 hospital beds per 1000 persons, which is much lower than countries such as South Korea (11.5) and the KSA, which has 2.7 (UN Report, 2020).

A close look at the extracted topics also highlights the focus of both governments on moving to digital platforms. This can be reflected through topic Ids 3 and 12 for the UAE, whereas for the KSA, the corresponding topic Ids are 10 and 13. Digital solutions and technology have played a significant role in providing essential services after the implementation of strict mitigation regulations as a result of the pandemic. The UAE' Ministry of Community Development (MOCD) has switched to using digital channels for government services. The UAE' Ministry of Education has implemented distance learning from March 2020 for all public/private schools as well as higher educational institutions, ensuring a safe and successful learning process. The KSA's government has also accelerated digital transition as a result of COVID-19. The KSA's government and private sectors developed and launched approximately 19 apps to manage public health services. Learning processes in Saudi Arabia also continued with the use of an established electronic learning infrastructure with a promising direction towards a wider adoption in the future (Hassounah et al., 2020).

Both the UAE and the KSA have also responded to air travel situation through different policy measures, as highlighted by topic Id 20 for the UAE and topic Id 4 for the KSA. Travel and tourism both constitute an important industry in the UAE, contributing 11.5 percent of its GDP. This industry was severely impacted in the wake of the pandemic restrictions on travel. Compared to 2019, the most considerable fall of scheduled departure flights in the UAE occurred on June 1, 2020, and equaled the decline of 82% (Aburumman, 2020). The UAE' government regularly updated the information for travellers on the Ministry's website as well as through other communication channels in terms of the restrictions and requirements for travelling to the UAE. Saudi Arabia took proactive measures to restrict the travel from China and other countries impacted by COVID-19 as early as February 2020, i.e. even before the first case was reported in March 2020.

In terms of differences in the topics extracted for two countries, there are few noticeable differences. There is a mention of artificial intelligence (AI) in one of the topics (topic Id 12) for the UAE, which could imply focus of the UAE' government on using AI for public policy. The UAE' government launched the 'UAE strategy for Artificial Intelligence', which played a vital role in containing the spread of COVID-19 in the UAE. For the KSA, there is a mention of 'Hajj' and 'pilgrim' in one of the topics (topic Id 7), indicating the KSA's government's response to the Hajj travel for 2020. Only around 100 pilgrims attended the Hajj in 2020 compared to the usual attendance of more than two million people. International travellers were restricted from the Hajj in 2020 and the worshippers were either foreign residents of Saudi Arabia or Saudi nationals (CNN, 2020). Another differing topic was 'banking related topic' (topic Id 15), which was present only for the UAE. The UAE has more banks and digital customers than the KSA. The UAE has around 50 banks in total, whereas the KSA has around 30 banks that are operational in the country.

Conclusion

Impacting the countries around the world, COVID-19 has caused the most severe pandemic of this century and has presented an uphill task for the governments to contain the pandemic and revive the economy through various public policies. In response to the pandemic, a large number of policy documents, announcements, updates, and guidelines have been published by governments. However, the huge volume of documents and publications makes it difficult to understand the key themes of government policies in an analytical way. By applying topic modelling and text clustering on those documents published by the UAE and the KSA which are related to government policies on COVID-19, we illustrated the fact that these techniques could help to analyse large volumes of data, as well as they could also facilitate the process of comparing key themes across countries through this analysis.

We used the LSA method for text mining and topic modelling. We compared the results for the UAE and the KSA by first comparing the key terms. We then uncovered hidden themes in the documents through text clustering and compared the clusters for the UAE and the KSA. We observed that the UAE and the KSA had similar themes in terms of government policy measures. For the UAE, the largest cluster is related to ‘Digital’ (31% of the total of the UAE’ documents), whereas for the KSA, the largest cluster is ‘International Travel’ (28% of the total KSA’s documents). Finally, we performed topic extraction analysis in order to identify key topics from the policy documents of the UAE and the KSA respectively. Most of the topics extracted for the UAE and the KSA have similarities between them. Some of the common similar topics revolve around ‘Vaccine Trial’, ‘Supporting Private Sector’, ‘Health Ministry’, ‘Digital Platforms’, and ‘Travel Restrictions’. Both governments announced and initiated various measures throughout 2020 in order to tackle COVID-19. There are few differences in topics for the UAE and the KSA. For the UAE, there are topics of ‘Artificial Intelligence’ and ‘Banking’, which are not found for the KSA. For the KSA, there is a mention of ‘Hajj’, which is due to the presence of Mecca and Medina in Saudi Arabia.

This study contributes to text-mining literature by providing a framework for analysing huge volumes of policy documents at the country level. Currently, there is a gap in research in terms of using topic modelling in general and LSA in particular for the text mining of policy documents. This paper can help to understand the key themes in the governments’ policies and can also help to identify success and failure of policy measures in certain cases by means of comparing the outcomes. In the era of digitisation and social networking, it is of immense importance to utilise advanced text mining algorithms to understand the themes in a fast and efficient way. This would enable governments and policy makers to adapt to the changes and tackle unprecedented challenges such as COVID-19 in an effective manner.

There are several limitations to this study. First, it does not cover all data related to the topic, as it is not possible to explore all the related articles because of time and access-rights issues. Second, we used one method of topic modelling only – namely the LSA – and we did not investigate how it compares with other methods. Third, the study does not take into consideration the difference in the quality of public policy documents of the UAE and the KSA, which may vary. Lastly, the study uses the SAS tool for the analysis, which is not a free software. However, the approach and methodology can be easily replicated using a free software such as R. In the future, this study could be enhanced by considering the most recent data and also extending its scope over social media data in order to validate the findings from this analysis. This study provides a bird’s-eye view of the response of the two countries to the COVID-19 crisis based on the text mining of public policy documents. In the future, the research hypothesis could be made more specific in order to address a particular section or topic for the analysis. The study can also be expanded by considering other approaches of topic modelling, in particular LDA, and then comparing the outcome with the present analysis.

References

Aburumman, A. (2020). The COVID-19 impact and survival strategy in business tourism market: The example of the UAE MICE industry. *Humanities & Social Sciences Communications*, available at: https://www.nature.com/articles/s41599-020-00630-8

Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, *6*(1), 147–153. https://doi.org/10.14569/ijacsa.2015.060121

Al-Obeidat, F., Kafeza, E., & Spencer, B. (2018). Opinions Sandbox: Turning Emotions on Topics into Actionable Analytics. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, *206*, 110–119. https://doi.org/10.1007/978-3-319-67837-5_11

Asmussen, C. B., & Møller, C. (2019). Smart literature review: A practical topi modelling approach to exploratory literature review. *Journal of Big Data*, *6*(1), 91, https://doi.org/10.1186/s40537-019-0255-7

Bechor, T., & Jung, B. (2019). Current State and Modeling of Research Topics in Cybersecurity and Data Science. *Journal of Systemics, Cybernetics and Informatics, 17*(1), 129–156.

Benedetto, F., & Tedeschi, A. (2016). Big data sentiment analysis for brand monitoring in social media streams by cloud computing. *Studies in Computational Intelligence, 639*, 341–377,https://doi.org/10.1007/978-3-319-30319-2_14

Carracedo, P., Puertas Medina, R., & Luisa Martí Selva, M. (2020). Research lines on the impact of the COVID-19 pandemic on business. A text mining analysis. *Journal of Business Research, 132,* 586–593. https://doi.org/10.1016/j.jbusres.2020.11.043

Cheng, X., Cao, Q., & Liao, S. S. (2020). An overview of literature on COVID-19, MERS and SARS: Using text mining and latent Dirichlet allocation. *Journal of Information Science*, September 2020. https://doi.org/10.1177/0165551520954674

CNN (2020). 'Unprecedented' Hajj begins – with 1,000 pilgrims, rather than the usual 2 million, https://edition.cnn.com/travel/article/hajj-2020-coronavirus-intl/index.html

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Goel, I., Sharma, S., & Kashiramka, S. (2021). Effects of the COVID-19 pandemic in India: An analysis of policy and technological interventions. *Health Policy and Technology*, *10*(1), 151–164. https://doi.org/10.1016/j.hlpt.2020.12.001

Gottipati, S., Shankararaman, V., & Lin, J. R. (2018). Text analytics approach to extract course improvement suggestions from students' feedback. *Research and Practice in Technology Enhanced Learning*, *RPTEL 13, 6.*https://doi.org/10.1186/s41039-018-0073-0

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76. https://doi.org/10.4304/jetwi.1.1.60-76

Hassounah, M., Rasheel, H. & Alhefzi, M. (2020). Digital Response During the COVID-19 Pandemic in Saudi Arabia. *Journal of Medical Internet Research, 22*(9), e19338

He, W., Tian, X., Hung, A., Akula, V., & Zhang, W. (2018). Measuring and comparing service quality metrics through social media analytics: A case study. *Information Systems and E-Business Management*, *16*(3), 579–600. https://doi.org/10.1007/s10257-017-0360-0

Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, *42*(1–2), 177–196. https://doi.org/10.1023/A:1007617005950

KPMG Report for Saudi Arabia, November 2020, available at: https://home.kpmg/xx/en/home/insights/2020/04/saudi-arabia-government-and-institution-measures-in-response-to-covid.html

Madhoushi, Z., Hamdan, A. R., & Zainudin, S. (2015). Sentiment analysis techniques in recent works. *Proceedings of the 2015 Science and Information Conference, SAI 2015*, *March*, 288–291. https://doi.org/10.1109/SAI.2015.7237157

Raja, A., Alshamsan A., & Al-Jedai, A. (2020). Current COVID-19 vaccine candidates: Implications in the Saudi population. *Saudi Pharmaceutical Journal, 28*(2020), 1743–1748

Samuel, J., Ali, G. G. M. N., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). COVID-19 public sentiment insights and machine learning for tweets classification. *Information (Switzerland)*, *11*(6), 1–22. https://doi.org/10.3390/info11060314

Sarkar, D., Bali, R., Sharma, T., Sarkar, D., Bali, R., & Sharma, T. (2018). Analyzing Movie Reviews Sentiment. *Practical Machine Learning with Python*. https://doi.org/10.1007/978-1-4842-3207-1_7

Sebei, H., Hadj Taieb, M. A., & Ben Aouicha, M. (2018). Review of social media analytics process and Big Data pipeline. *Social Network Analysis and Mining*, *8*(1), 1–28. https://doi.org/10.1007/s13278-018-0507-0

Sharma, A., Adhikary, A., & Bikash, S. (2020). Covid-19's impact on supply chain decisions: Strategic insights from NASDAQ 100 firms using Twitter data. *Journal of Business Research*, *117*(May), 443–449. https://doi.org/10.1016/j.jbusres.2020.05.035

Shi, L., Tsai, J., & Kao, S. (2009). Public health, social determinants of health, and public policy. *Journal of Medical Science, 29*(2), 43–59.

United Nations COVID-19 Socio-Economic Analysis for the United Arab Emirates, UN Report, September 2020.

Walker, R. M., Chandra, Y., Zhang, J., & van Witteloostuijn, A. (2019). Topic Modeling the Research-Practice Gap in Public Administration. *Public Administration Review*, *79*(6), 931–937. https://doi.org/10.1111/puar.13095

Wesslen, R. (2018). Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond. *ArXiv*

Xu, J., Tao, Y., Yan, Y., & Lin, H. (2018). VAUT: A visual analytics system of spatiotemporal urban topics in reviews. *Journal of Visualization*, *21*(3), 471–484. https://doi.org/10.1007/s12650-017-0464-0

Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. *Complex & Intelligent Systems*, *6*(3), 621–634. https://doi.org/10.1007/s40747-020-00155-240747-020-00155-2

Appendix

Various government entities and media reports citing government actions as the COVID-19 response have been taken from the following websites:

- UAE Embassy https://www.uae-embassy.org/
- Federal Authority for Identity and Citizenship (ICA) https://smartservices.ica.gov.ae/
- Ministry of Foreign Affairs https://www.mofaic.gov.ae/
- News and Media: https://www.khaleejtimes.com/
- News and Media: https://english.alarabiya.net
- News and Media: https://gulfnews.com
- News and Media: https://www.meed.com/
- News and Media https://www.arabnews.com/
- News and Media https://www.reuters.com/
- Ministry of Health and Preventions: https://www.mohap.gov.ae/
- News and Media: http://wam.ae/
- News and Media: https://www.jdsupra.com/
- News and Media: https://www.aljazeera.com/
- National emergency crisis and disaster recovery: https://www.ncema.gov.ae/
- Privately owned security services company: https://www.garda.com/
- Community platform for real estate: https://bldgtmrw.com/
- News and Media: https://www.cnbc.com/
- The National Emergency Crisis and Disasters Management Authority's platform: http://www.weqaya.ae/
- News and Media: https://www.caixinglobal.com/
- Ministry of Health KSA initiative: https://covid19awareness.sa
- News and Media: https://www.bbc.com/
- International Monetary Fund: https://www.imf.org/
- Ministry of Health KSA: https://www.moh.gov.sa/
- Saudi Arabia Monetary Authority: http://www.sama.gov.sa/
- News and Media: https://www.atlas-mag.net/
- News and Media: https://thearabweekly.com/
- The Saudi Data and Artificial Intelligence Authority: https://sdaia.gov.sa/
- Integrated encyclopedia: https://mhtwyat.com/
- Johns Hopkins Aramco healthcare: https://www.jhah.com/
- Saudi Press Agency: https://www.spa.gov.sa/