Social Network Analysis and Mining (2022) 12:26
https://doi.org/10.1007/s13278-022-00857-0

ORIGINAL ARTICLE

A comparative analysis of the world's constitutions: a text mining approach

Tuncay Bayrak[1]

Received: 3 November 2021 / Revised: 3 January 2022 / Accepted: 5 January 2022

Abstract
A constitution is the foundation for a government in almost every society around the world. Analyzing and studying constitutions of different countries may allow one to understand better how various countries form the basic structure of their governments, the legal and cultural aspects of their people, the responsibilities of key institutions, and the most basic rights of the people. In this study, we performed text mining to make a comparison between the US constitution and the constitutions of various countries located in four different geographic regions of Europe, Asia, the Middle East, and Africa with respect to how they define and outline the liberties and rights of their citizens and the relationship between the key institutions of their governments.

Keywords Constitution analysis · Text mining · Textual data

1 Introduction

A constitution describes the basic structure and the foundation of any government. Through their constitutions, governments define, describe, and outline the rights of their citizens, specify the relationship between the key institutions of government, describe how the powers of the government separated into different branches, and describe how the branches work together. They furthermore outline the duties of the governments toward their citizens. Simply put, a constitution defines the principal organs of government and their jurisdictions and the basic rights of citizens (Britannica 2021).

While the content and nature of a particular constitution varies considerably between countries, the vast majority of contemporary constitutions describe the basic principles of the state, the structures and processes of government and the fundamental rights of citizens in a higher law (Bulmer 2017). One might argue that a typical constitution in a democratic nation serves as a blueprint for how different branches of a government operate and interact with one another.

Since manually analyzing large collections of textual data such as constitutions is no longer effective, qualitative data analysis techniques such as text mining may aid one in understanding how various countries in their constitutions define the basic structure of their governments, how the liberties and rights of the people are defined and described, and how the power distribution and the responsibilities of key institutions are specified.

We took a text mining approach to make a comparison between the USA constitution and the constitutions of various countries located in four different geographic regions of Europe, Asia, the Middle East, and Africa with respect to how they define and outline the liberties and rights of their citizens and the relationship between the key institutions of their governments. The Spanish speaking Latin American countries are excluded from this study as they were previously surveyed and analyzed in a different study (Wolf 2014). In this study, we employed SAS Enterprise Miner, a text mining and analytics tool, to survey whether the constitutions of such counties in different geographic regions allude to the same core principles. Our aim is to provide a systematic investigation of constitutions of a number of countries in an attempt to identify patterns of differences and similarities between these constitutions.

✉ Tuncay Bayrak
tbayrak@wne.edu

[1] College of Business, Western New England University, 1215 Wilbraham Rd., Springfield, MA 01119, USA

Published online: 18 January 2022

2 Relevant literature

Using text mining tools, textual data contained in documents such as constitutions can be analyzed to transform unstructured data into structured data so that patterns and themes can be identified, and valuable insights can be extracted.

The amount of data captured and stored has been enormously increasing day by day which is generally in the unstructured form (Akilan 2015). Although textual data are the easiest form of unstructured data which can be created in any application scenario, there has been a tremendous need to effectively process a wide variety of textual data (Aggarwal and Zhai 2012). In addition, since an unprecedented expansion in the volume of unstructured data in digital textual formats has been witnessed in recent decades, extracting meaningful and relevant information from the voluminous unstructured data remains a challenge (Ittoo et al. 2016). Therefore, analyzing unstructured textual data such as constitutions calls for employing text mining or text analytics tools. In other words, the proliferation of digital textual data magnified the need for the inventions of efficient ways of extracting information from the textual data sources (Kwayu et al. 2021). Through text mining applications and tools, vast collections of textual data can be processed and analyzed to capture key themes, patterns, concepts, trends, and hidden relationships. As pointed out by Ittoo et al. (2016), to recognize the potential economic value lying untapped in their text data repositories and sources, companies can employ text mining and analytics techniques.

Text mining may be defined as the process of extracting meaningful information and discovering different interesting text patterns (Akilan 2015; Aggarwal and Zhai 2012). From a slightly different perspective, IBM (2020) defines text mining as "*the process of transforming unstructured textual data into a structured format to identify meaningful patterns and new insights*." Kobayashi et al. (2018) argue that text mining involves a quantitative approach to the analysis of voluminous textual data and helps accelerate knowledge discovery by radically increasing the amount of data that can be analyzed.

Ittoo et al. (2016) suggest that information extracted from textual data sources may be valuable for data-driven and informed decision-making. Moreover, text mining unstructured textual data may help to discover and understand the hidden content behind the large collections of unstructured bodies of text (Ibrahima and Wang 2019) and provide a valuable source of business insights (Gruss et al. 2018).

Since text mining techniques may allow one to uncover hidden relationships in textual data previously not available and gain insights into key concepts, trends, and patterns, they have been employed in various domains and fields by researchers. For instance, Hwang et al. (2020) used text mining techniques to examine police-recorded data for 1,601 domestic violence events involving autism and 414,840 events not involving autism in the state of New South Wales, Australia from January 2005 to December 2016. Using consumer reviews data extracted from Expedia.com, Xiang et al. (2015) employed text mining techniques to analyze customer satisfaction with services provided by a hotel. A text mining study conducted by Ibrahima and Wang (2019) surveyed customers' topics of concern that are shared among Twitter users regarding online retail brands. Moreover, a text mining approach was applied by Al-Daihani and Abraham (2018) to a dataset collected from the Facebook posts of academic libraries at the top 100 English-speaking universities. Chen et al. (2020) performed text mining to explore the mechanism of doctor-patient interaction in an online health community.

The literature on text mining suggests that various text mining techniques and methods may be employed to develop classification and predictive models as well. For instance, Xu et al. (2021) used a text mining technology to develop an improved approach to identify and predict safety risk factors from a volume of construction accident reports. Using text mining tools, Torii et al. (2015) developed detection systems to identify relevant risk factors in electronic medical records. Employing a similar method, Zhang et al. (2020) developed text mining and machine learning techniques capable of correctly classifying technician reports. Chatterjee et al. (2021) utilized text mining techniques to predict customer satisfaction with core service aspects in the health care industry. Text mining applications were also employed in the past to analyze and predict insurance fraud (Wand and Xu 2018). Recently, Sohrabi et al. (2020) introduced a model to predict tourists' destinations based on their interests and travel backgrounds.

By effectively gathering and analyzing textual data, companies furthermore can develop highly competitive products (Bigorra et al. 2020), facilitate organizational learning and enhance citizen-centric public service quality (Reddick et al. (2017), manage performance of companies (Kang et al. 2018), and measure dimensions of employee-visitor encounters that may have an impact on visitor outcomes.

Since advanced text mining techniques were not available in the 80s and 90s, several studies conducted in the 80s and 90s analyzed the US constitution using traditional research method practices and data analysis techniques. For instance, Galloway (1988) performed an analysis on the US constitutions with respect to the proper functioning of government and the protection of individual liberties. Killian and Costello (1992) carried out a similar study to provide a more detailed analysis of the US constitution. Their analysis notes

the rise of the equal protection clause as a central concept of constitutional jurisprudence. A year later, Boudreaux and Pritchard (1993) developed an economic theory of the constitutional amendment process focusing particularly on the roles that Congress and interest groups play in that process. The authors conclude that while the founders intended to put the Constitution beyond the reach of factions, that intend was not realized. The authors further argue that although factions cannot control the content of the Constitution, neither can the majority. And finally, the role of tradition in constitutional interpretation was investigated in a study conducted by Pritchard and Zywicki (1998). The authors hold that constitutionally significant traditions can be identified.

A couple of more recent studies leveraged text mining tools to examine the constitutions of the Spanish speaking South American countries (Wolf 2014), and Brazilian and Portuguese constitutions (Belokurows, n.d). Wolf (2014) found that the Spanish term "derecho," or "right" as in your "right" as a citizen is used more frequently than any other terms in Spanish speaking South American countries' constitutions. Additionally, the majority of them have the term "president" in their constitutions as well as consider themselves republics as the term "república" appears quite frequently in their constitutions.

While the vast majority of contemporary constitutions describe the basic principles of the state, the structures and processes of government and the fundamental rights of citizens (Bulmer 2017), analyzing them through text mining may provide a deeper understanding of the similarities and differences of the constitutions of different countries. To the best of our knowledge, since no study in the past has employed text mining tools to make a comparison among the constitutions of various nations located in different geographic regions with different cultures, languages, and government structures, in the following sections, we took a text mining approach to examine the similarities and differences within these constitutions.

3 Research method

To perform text mining and make a comparison between the constitutions, we first picked twenty countries in each region. The countries selected for this analysis are listed in Table 1. Although some counties such as Pakistan and Indonesia are geographically located in Asia, we put them together into the same group as the Middle Eastern countries as they share the same religion. We wanted to see whether the constitutions of the predominantly Muslim nations share any similarities with the constitutions of the countries located in other geographic regions.

Having identified the constitutions included in this study, we retrieved the HTML versions of each constitution from the website www.constituteproject.org, which provides access to the full text of the world's constitutions in English, and classified the constitutions into five main categories of USA, Europe, Asia, the Middle East, and Africa and performed text mining on them separately using the text mining and analysis tools available in SAS Enterprise Miner, a powerful commercial business analytics application. The point of text mining is to attempt to extract meaningful information from unstructured textual data. This can be done systematically by a software application which automatically breaks down trends and patterns in order to draw insightful conclusions. Unlike the typical numerical data, text mining draws from linguistic variables to better understand what is contained in a document or a text corpus.

Table 1 Countries selected for this analysis

North America	Europe	Asia	Middle East	Africa
USA	Germany	China	Morocco	South Africa
	France	Japan	Egypt	Namibia
	Italy	Vietnam	Turkey	Angola
	Spain	Philippines	Libya	Zimbabwe
	Portugal	Georgia	Tunisia	Botswana
	Greece	Malaysia	Algeria	Kenya
	Austria	India	Iran	Ethiopia
	Belgium	Russia	Iraq	Chad
	Denmark	Mongolia	Jordan	Congo
	Netherlands	Cambodia	Yemen	Cameroon
	Poland	Thailand	Saudi Arabia	Zambia
	Bulgaria	Nepal	Bahrain	Uganda
	Romania	Tajikistan	Kuwait	Tanzania
	Hungry	Kyrgyzstan	Syria	Nigeria
	Ukraine	Uzbekistan	Oman	Niger
	Serbia	Kazakhstan	Qatar	Ghana
	Norway	Turkmenistan	Lebanon	Senegal
	Sweden	Myanmar	Pakistan	Gabon
	Finland	Bhutan	Bangladesh	Mozambique
	Switzerland	Singapore	Indonesia	Madagascar

Having uploaded the text file into SAS Enterprise Miner to run text mining, the first node we used was the Text Parsing node, which is used to deconstruct and break down the text file. Next, we employed the Text Filter node to filter out the extraneous information and terms in the text file. Filtering helps to drop any terms that are not relevant for analysis and are nonessential for developing meaningful conclusions about the text file.

The final nodes are the Text Cluster and Text Topic nodes, which are used to cluster the remaining terms from the text file after they have been parsed and filtered, and to define combinations of words or collections of terms that

describe and characterize a main theme or idea in a document collection.

The text mining and analysis process we employed in this study is depicted in Fig. 1. A detailed description of each step is given in the following sections.

In the following section, we provide a systematic and comprehensive analysis of the constitutions of the countries in five different geographic regions.

4 Data analysis and results

4.1 Text parsing and analysis

Once we compiled the constitutions into text files and upload them to SAS, we employed the Text Parsing node in SAS Enterprise Miner to break down and deconstruct the constitutions into terms and to identify and quantify information about the terms contained in the constitutions. In other words, the Text Parsing node is often utilized when analyzing unstructured textual data to gather statistical data about the terms such as the frequency of occurrence of each term in a document collection. Table 2 tabulates and summarizes the most frequently appearing terms in the constitutions of the countries across five different regions. For instance, the term "law" appears 4367 times throughout the constitutions of twenty different countries located in Europe. As expected, such important terms as "state," "law," and "president" are among the most frequently appearing terms in the constitutions across five different regions.

Using the data tabulated in Table 2 as our data source, we created a pivot chart to help visualize the distribution of the most common terms (Fig. 2). Browsing through the regions and the terms, a number of interesting observations can be made. For instance, certain terms such as "united," "congress," house," and "power" are among the most frequently appearing terms in the US constitution. However, they are not as common in the constitutions of the countries in other four regions. Similarly, while the terms "council," "public," and republic" are among the most common terms in the constitutions of the Middle Eastern countries, no other regions' constitutions mention them as frequently. In addition, the terms "federal," "provision," and "matter" are cited a lot more frequently in the constitutions of the European counties than in that of other regions. In their constitutions, African countries cite the term "person" more frequently than any other four regions. Finally, while the term "assembly" appears 2221 times in the constitutions of African countries, it is not among the most frequently cited terms in other four regions. These differences may be attributed to the fact that different countries have different government structures and processes.

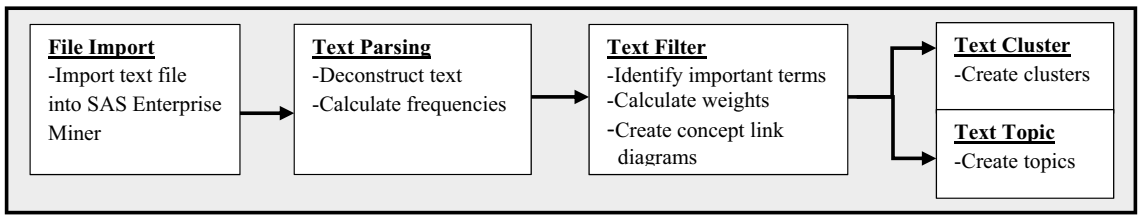

Fig. 1 Text mining and analysis process using SAS Enterprise Miner

Table 2 Top 10 most frequently appearing terms in each region

USA		Europe		Asia		Middle East		Africa	
Term	Freq	Term	Freq	Term	Freq	Term	Freq	Term	Freq
state	217	law	4367	state	3337	law	2991	person	3885
president	121	member	2275	law	3315	member	1662	law	3817
united	88	right	1980	member	1955	state	1583	president	3418
congress	62	government	1921	person	1764	president	1395	office	3240
office	57	federal	1824	president	1702	right	1226	member	3212
law	52	state	1548	constitution	1633	constitution	1104	constitution	3192
person	49	provision	1394	government	1323	government	1019	parliament	2784
house	43	national	1319	right	1440	council	988	state	2727
power	37	president	1258	parliament	1247	public	956	national	2622
constitution	32	matter	1253	office	1241	republic	877	assembly	2221

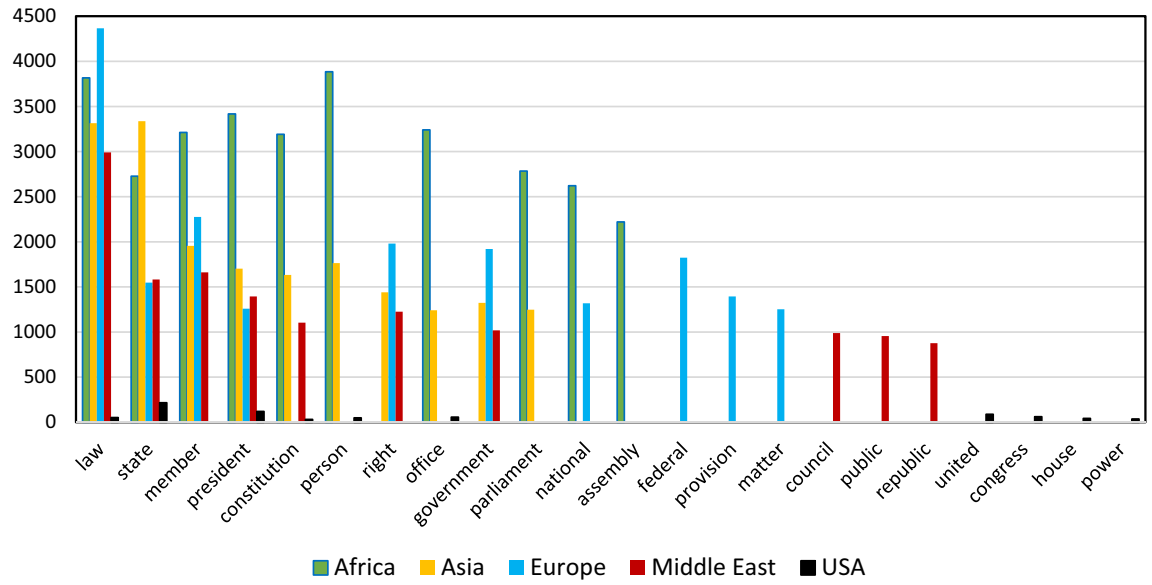

Fig. 2 Most common terms by regions

4.2 Text filtering and identifying significant terms

The Text Parsing node allows one to explore the terms in a document collection and generate simple statistical data about the terms. However, it does not allow us to identify and capture the most relevant terms and information. Therefore, to identify the most important terms in any given document collection and eliminate extraneous information, the text filter node is employed when performing text mining on unstructured textual data.

While several term weighting methods are available, the default term weighting method employed in SAS Enterprise Miner is Entropy. With the entropy method, the weight for term *wi* is computed using the following formula:

$$wi = 1 + \sum_{j=1}^{n} \frac{(\frac{fij}{gi}).log_2(\frac{fij}{gi})}{log_2(n)}$$

where g_i is the number of times that *ith term* appears in the document collection, n is the number of documents in the collection, and f_{ij} is the *ijth* frequency in the unweighted term-by-document matrix (Chakraborty et al. 2013). The entropy method assumes that terms that are used to categorize documents are those that occur in only a few documents but they occur many times in those few documents, and those terms are significant in discriminating the documents (Chakraborty et al. 2013).

Table 3 tabulates the most important terms found in the constitutions of the countries surveyed in this study. As seen, having parsed the constitutions and eliminated extraneous

Table 3 Most important terms and their weights

USA		Europe		Asia		Middle East		Africa	
Term	Weight	Term	Weight	Term	Weight	Term	Weight	Term	Weight
act	0.768	intergovernmental	0.877	self-governing body	0.884	benches	0.879	pillar	0.958
vacancy	0.761	professor	0.872	consignment	0.878	bar	0.879	watershed	0.896
choice	0.761	judge	0.872	lieutenant	0.875	judicial office	0.870	lake	0.888
death	0.759	forest	0.868	prince	0.875	zone	0.870	bearing	0.886
money	0.759	budget management	0.863	food	0.875	food	0.870	edge	0.883
high	0.759	national asset	0.863	fire	0.875	allegation	0.870	east	0.883
appointment	0.759	inspector	0.863	bail	0.875	fiqh	0.870	rock	0.880
treason	0.759	partner	0.863	palace	0.875	port	0.867	undistributed	0.880
first	0.759	electoral register	0.863	minimum standard	0.866	regent	0.861	distance	0.875
present	0.759	whole ordinance	0.863	written declaration	0.866	cadre	0.861	highest point	0.873

information, a different set of terms emerged. The Text Filter node ranks the terms based on their weights or importance. In other words, any term with a value for weight greater than zero is considered important, and all terms with weight equal 0 do not have much significance. The value of weight determines the terms' importance. Therefore, a large weight represents a term has higher degree of importance and a lower weight represents a term is relatively less important.

As seen in Table 3, the term "act" holds the highest significance for the US constitution. Similarly, the term "intergovernmental" bears the highest importance for the European constitutions. With respect to Asian countries, they seem to place the highest weight on the term "self-governing body." While the Middle Eastern countries assign the highest weight to the term "bench", the term "pillar" assumes the highest weight in the constitutions of African nations.

As summarized in Table 2, while some of the most frequently appearing terms such as "law" are found in different constitutions of the countries located in five different regions, almost none of the terms tabulated in Table 3 appear across different constitutions of various countries located in five different geographic regions. Scanning through the terms in Table 3, we came across one particular term "*fiqh*" that is exclusively used in the constitutions of the Middle Eastern countries. The literature suggests that it may be translated as "Islamic jurisprudence." It appears, given the fact that Middle Eastern nations are predominantly Muslim nations, it is only natural for them to make a reference to the term "Islamic jurisprudence" in their constitutions.

4.3 Relationship among terms—concept link

We carried out a further analysis to examine the most important terms to see how they are related to and associated with other terms in the constitutions. In unstructured textual data analysis, concept link diagrams aid one in understanding the relationships between various terms based on the co-occurrences of the terms in the same documents (Chakraborty et al. 2013). In other words, they depict and illustrate how the most important terms in a document collection are related to one another. The term with the highest weight or the selected term appears at the center of the diagram, surrounded by the terms that are most highly associated with the term. The width of the link in the diagram shows the strength of the association between the terms. Concept linking diagrams also allows one to understand the context in which a particular term is frequently used (Chakraborty et al. 2013).

As tabulated in Table 3, the term "act" attained the highest weight or importance in the US constitution. The concept linking diagram in Fig. 2 shows a hyperbolic tree graph with the term "act" in the center of the tree structure, and the other terms that are strongly associated with it. As seen, the term "act" is closely associated with the terms "president," and "elect." We further expanded the diagram to see how those two terms are associated with some other terms. As seen, not surprisingly, the term "president' is strongly associated with such terms as "senate," "vice," "transmit," and "representative." Similarly, the term "elect" is closely related to the terms "office," "person," "term," and "choose."

The US constitution speaks of who will *act* as the president in case of the removal of the president from office, or of his death, resignation, or inability to discharge the powers and duties of the said office. Figure 3 also suggests that a strong relationship exists between "president" and "vice president," "senate," and "representatives." In addition, some other important terms such as "term," "year," and "office," pertaining to "election" are depicted in Fig. 3. It may be argued that the terms in Fig. 3 depict the relationships among the various elements of the legislative branch of the US government. Surprisingly, Fig. 3 makes no reference to the executive and judicial branches of the US government.

A similar analysis was performed on the constitutions of the European countries. As illustrated in Fig. 4, a strong association exists between the terms "intergovernmental" and a number of other important terms such as "frame," "foreign state," "transfer," and "state treaty." Having further expanded the term "frame", we obtained several important terms such as "legal," "union," "federal," and "national."

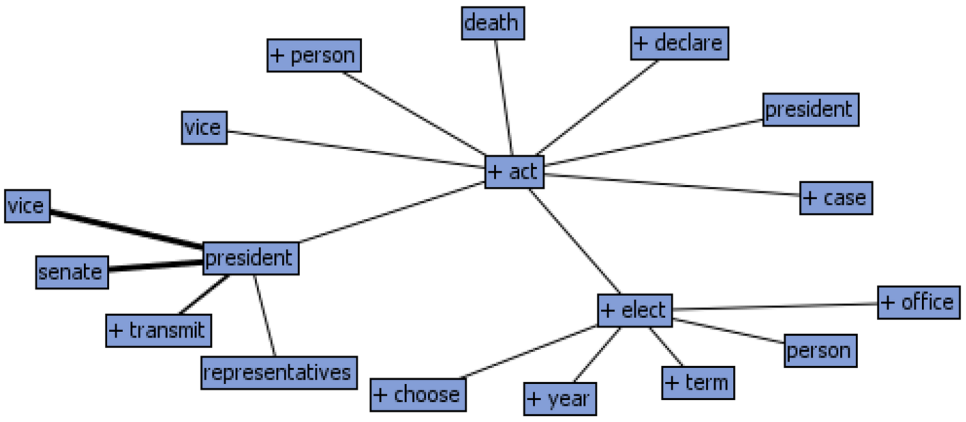

Fig. 3 USA constitution: concept link

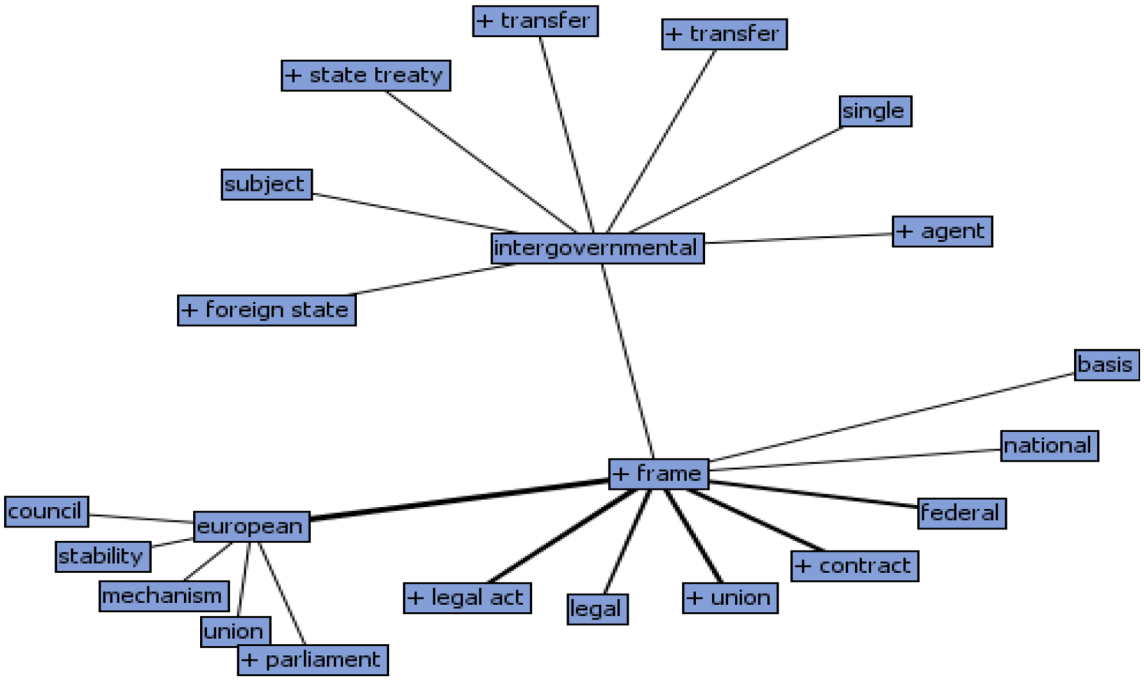

Fig. 4 European constitutions: concept link

The concept link diagram in Fig. 4 implies that European countries in general appear to see their constitutions as a "basis" or a "frame" and "legal contract" for their "national" and "federal" "union." Moreover, their constitutions seem to address a set of important issues pertaining to the European Union's stability, parliament, and council. European constitutions also speak of "transfer" of existing administrative institution and sovereign powers to international organizations, and the responsibilities and powers of the Federal Banks to the European Central Bank. Finally, in their constitutions, they refer to how the Constitutional Court pronounces whether "state treaties" signed with "foreign states" are contrary to law.

In regard to the Asian constitutions, the term "self-governing body" bears the highest weight (Fig. 5), which is strongly associated with the term "self-governing." Looking at the concept link diagram depicted in Fig. 5, it may be argued that while in their constitutions Asian countries draw attention to "state authority," they, on the other hand, seem to exercise their "state authority" through "self-governing" and "local administrative units" and the "decisions" made by these local units.

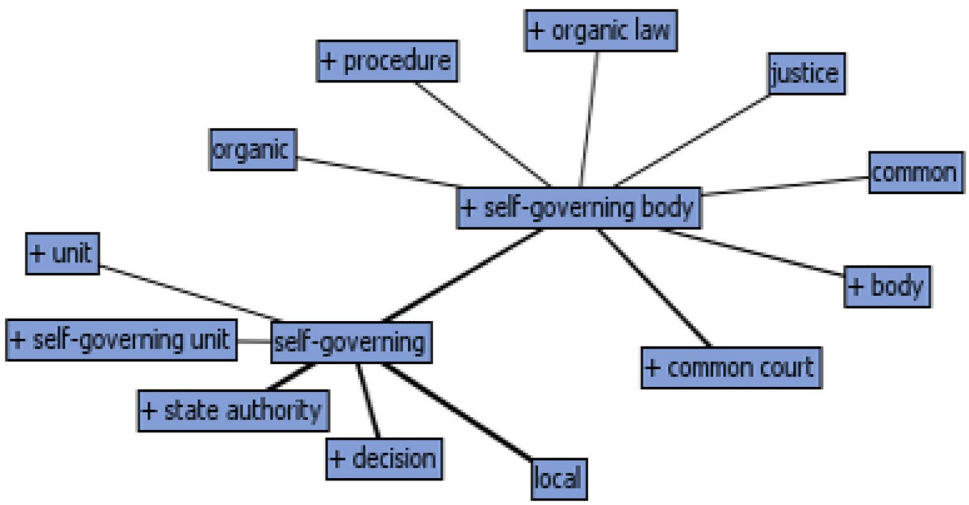

Fig. 5 Asian constitutions: concept link

Asian countries in their constitutions see a strong relationship between "organic laws" that form the foundation of their governments and their "judicial" and "court" system. Figure 5 further suggests that they assign a strong relationship between "self-governing body" and the procedures established by law. For instance, they speak of the procedure and the organic law for the election of Parliament, and how the Parliament fulfils full administrative functions in the House of Parliament in accordance with the procedures established by the Rules of Procedure.

The concept link diagram generated for the Middle Eastern countries' constitutions seems to portray a slightly different picture (Fig. 6). While the concept link diagram suggests a strong association between the terms "benches" and "treasury" and includes some important terms such as "law," "state," and "constitute," the Middle Eastern countries' constitutions highlight the importance of the function of "treasury" and "budget" managements for the "state" as no government can properly function without a solid budged and treasury department. While one would agree that the role of the treasury department is expected to be acknowledged by every country, it seems it is more pronounced in the constitutions of the Middle Eastern countries.

With respect to African nations, Fig. 7 suggests that African nations perceive their constitutions as the "pillar" of their "state," "unity," and "democracy." Expanding each important term yielded a set of more important terms. For instance, as expected, African nations' constitutions describe democratic principles and values, define citizens' rights, value pluralism, and the rule of law. In addition, Fig. 7 further suggests that African nations perceive their constitutions as a basis for "promoting" their "national unity" "independence" and "sovereignty," "peoples' integrity," "peace," and "democracy." Looking at Fig. 7, it may be suggested that democratic principles and values such as "pluralism" and citizens' "rights" form the foundations of their "societies."

The concept link diagram analysis of the world's constitutions conducted in this study reveals that countries in

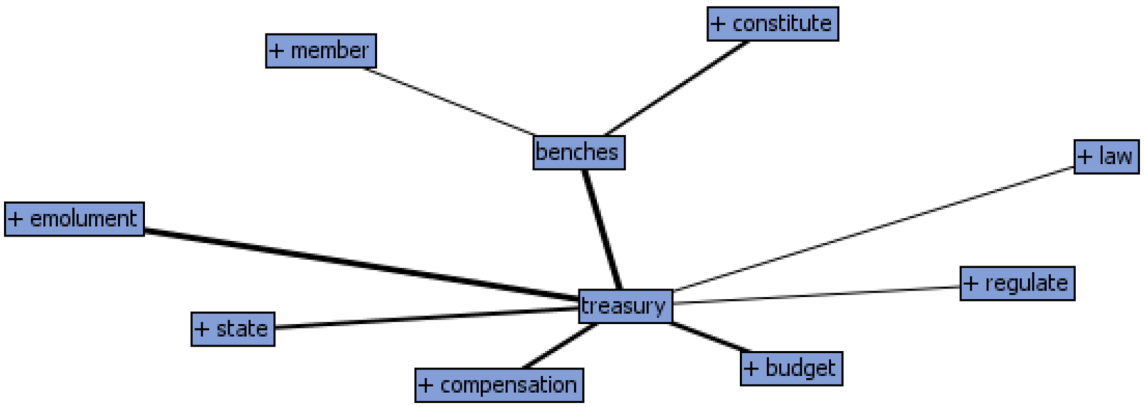

Fig. 6 Middle Eastern constitutions: concept link

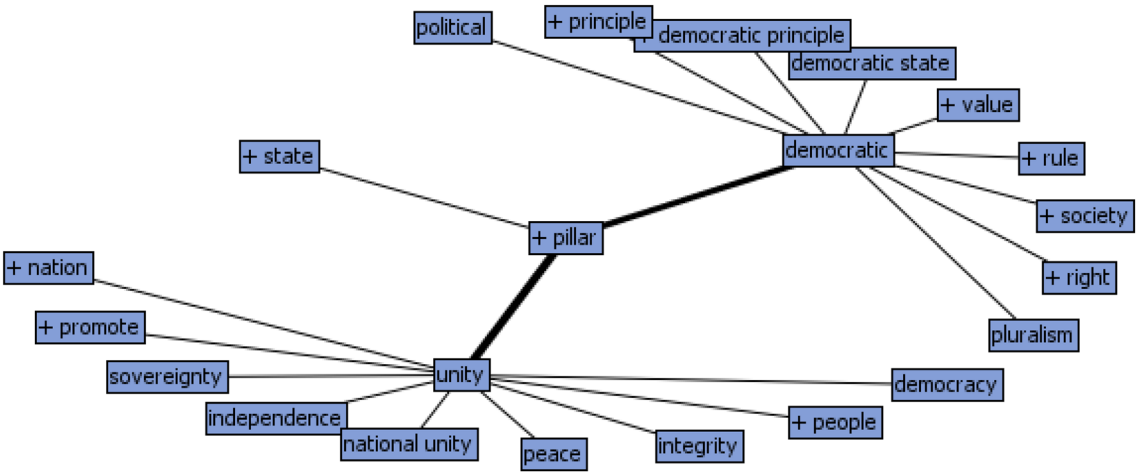

Fig. 7 African constitutions: concept link

every region make references in their constitutions to the legislative, executive, and judicial branches of government, the structure of the government, and the liberties and the rights of their citizens. However, some regions highlight and pronounce more the importance of the treasury and budged management (Middle Eastern countries), the role played by the Supreme Court in whether state treaties signed with foreign states are contrary to law (European countries), how the state authority exercises it power through local administrative units (Asian countries), and the importance of the office held by the president (USA).

4.4 Text mining and cluster analysis

To explore some similarities and differences of the constitutions and to identify themes within them and assign each term to one of the themes, we ran a cluster analysis using the Text Cluster node in SAS Enterprise Miner which splits the document containing the constitutions into non-overlapping clusters and uses a set of descriptive terms that best describe the theme of each cluster.

In the context of textual data, cluster analysis is often performed to discover groups of similar terms in the text corpus. In other words, it aims to discover natural groupings of terms and to present an overview of classes of terms in a collection of documents (Andrew and Fox 2007). In essence, cluster analysis divides the document collection into mutually exclusive groups based on the presence of similar themes (Chakraborthy 2013). By looking at the clusters, insights on the constitutions at a very high level can be gained.

Technically speaking, in text mining, cluster analysis involves natural language processing (NLP) algorithm that partitions and categorizes terms and words, grouping them into subsets or clusters. It creates clusters that are coherent internally, but clearly different from each other. In other words, documents within a cluster should be as similar as possible, and documents in one cluster should be as dissimilar as possible from documents in other clusters (Manning et al. 2008). Hence, the goal of a cluster analysis is to create the smallest number of distinct clusters with highly similar terms in the same cluster.

In this analysis, we utilized the hierarchical clustering algorithm where we set the number of clusters to be generated to 5 and the number of descriptive terms within each cluster to 10.

The result of a cluster analysis performed on the US constitution is summarized in Table 4, which shows cluster ID for each cluster, descriptive terms within each cluster, a percentage of the number of observations that got put into each cluster, and the clusters generated. As seen, cluster 6 is the largest cluster with such important terms as "senate," "president," "right," and "vote." The second largest cluster (cluster 11) has a similar content with terms like "office," "state," "president," and "senate." These two clusters essentially address the legislative and executive branches of the US government.

We took a similar approach to identify themes within the constitutions of European countries. Looking at Table 5, it may be suggested that while clusters 6 and 9 mention important terms such as "legal authority," "law," "court," and "constitution," that mainly addresses the judicial branch of government, cluster 3 speaks of the structure of government, and cluster 8 speaks of citizens' rights.

As for the constitutions of Asian countries, as seen in Table 6, the largest cluster (44%) includes important terms such as "parliament," "supreme court," "house," and "prime minister," suggesting that it addresses both the legislative and executive branch of typical government. The other three clusters with equal percentages

Table 4 USA constitution: text cluster

Cluster ID	Descriptive terms	%
6	constitution years senate president ratified seven + right + vote date fill	0.53
7	legislation 'appropriate legislation' enforce + place + time government + punishment + seat cases convicted	0.09
11	office law united duties states laws president constitution senate vice	0.21
12	persons members + person + case + right + law public electors + time vice	0.18

Table 5 European Countries constitutions: text cluster

Cluster ID	Descriptive terms	%
3	national + government council public members + office assembly republic + president matters	0.19
6	public + section local + provision court down constitutional + authority legal	0.49
8	+ right rights + person members republic + committee + constitution citizens + election	0.17
9	law federal + force + court constitutional administrative courts law court	0.15

address other basic elements of government such as the structure of the government (republic and federation) and persons' rights (cluster 5 and 9), national government and citizens' rights (cluster 7).

Performing a cluster analysis on the constitutions of the Middle Eastern countries resulted in 5 distinct clusters. The largest cluster (cluster 9, 47%) makes references to the judicial branch of the government (supreme court), and some functions and responsibilities of the treasury department using terms such as "financial," and "expenditure." Clusters 11 and 12 with the same weights address both the legislative and executive branches of a government as they include such terms as parliament, house, president, ministers, council, law, and court. Finally, cluster 7 deals with education institutions and citizens' rights (Table 7).

A cluster analysis performed on the constitutions of African nations yielded the clusters summarized in Table 8. As seen, cluster 5 (51%) seems to address the judicial branch of government (court, constitution, law) and citizens' rights and freedoms. Clusters 6 and 7 together speak of the structure of government (republic, parliament, national assembly) and social and economic services provided for their citizens.

4.5 Text topics

The Text Topic node identifies combinations of words or collections of terms that describe and characterize a main theme in a document collection. Topics are formed based on the fact that these terms frequently associate together. Terms can belong to more than one topic or to none at all. Essentially, similar terms are combined or merged into the same topic of interest. The approach is different from clustering because clustering assigns each document to a unique group. However, the Text Topic node assigns a score for each document and term to each topic. The scores are assigned by a matrix factorization technique known as singular value decomposition (SVD). Then, thresholds are used to determine whether the association is strong enough to consider that the document or term belongs to the topic (SAS 2021).

For each region, we wanted to create no more than five topics with a maximum of five terms that describe a main theme within each topic. By combining individual terms into topics, we can get a better understanding of the content of the constitutions of the countries located in different regions.

Through groupings of individual words and terms, the Text Topic node generated five topics for each region tabulated and summarized in Table 9. Each topic is made up of a number of similar words that describe a main theme.

Table 6 Asian Countries constitutions: text cluster

Cluster ID	Descriptive terms	%
5	+ law + right + president + person + court rights federation + order + case	0.19
6	+ parliament + constitution court + house + government supreme + number prime minister total	0.44
7	national + government + person assembly + people + citizen + development citizenship + clause years	0.19
9	+ state + law + office + president + act + republic + union + term + election act	0.19

Table 7 Middle Eastern Countries constitutions: text cluster

Cluster ID	Descriptive terms	%
7	+ state + right rights public political citizens social + protection education institutions	0.14
8	members national assembly + election first elections years elected + majority + period	0.11
9	court constitutional + expenditure consolidated supreme financial fund Allah act	0.47
11	+ government + constitution + court parliament court + house provisions federal representatives + person	0.14
12	+ law + council + people national + assembly + president + constitution provisions laws ministers	0.14

Table 8 African countries constitutions: text cluster

Cluster ID	Descriptive terms	%
4	+ office + person functions + election president + commission + power powers judicial years	0.14
5	+ law + state + court court + right rights political constitutional + council + freedom	0.51
6	+ government + president republic + constitution provisions + section local social economic + service	0.18
7	national assembly public + parliament members + law commission + state persons service	0.18

Table 9 Text Topics

Region	Topic ID	Topic	# of terms
USA	1	constitution, + ratify, several, + legislature, seven	15
	2	president, president, + office, vice, + house	22
	3	appropriate legislation, legislation, enforce, + power, congress	6
	4	+ law, + state, public, united, cases	19
	5	+ citizen, + number, + deny, + abridge, + vote	16
Europe	1	+ member, national, + government, assembly, council	228
	2	+ right, + law, public, + authority, + freedom	321
	3	federal, council, federal, law	155
	4	+ provision, + supplementary provision, supplementary, + law	117
	5	president, court, constitutional, republic	188
Asia	1	+ member, + house, assembly, national, + office	194
	2	+ right, federation, + citizen, + government, + state	252
	3	court, + judge, + court, supreme, + justice	161
	4	+ region, + state, + representative	132
	5	+ law, + constitution, + provision, + parliament, + state	226
Middle East	1	+ election, assembly, + member, + office, + day	184
	2	+ right, + state, + freedom, + citizen, + guarantee	211
	3	+ court, court, + judge, supreme, + office	158
	4	+ council, + government, minister, + minister, prime	169
	5	+ law, + provision, + constitution, parliament, + authority	219
Africa	1	national, assembly, + member, + parliament, + president	185
	2	court, + court, high, appeal, + justice	145
	3	+ parliament, + constitution, + law, + person, + provision	183
	4	+ office, + president, + person, + appoint, + hold	176
	5	+ state, + government, national, republic, + law	260

For instance, looking at the European constitutions, topic 2, which is made up of 321 terms, contains terms and words that describe both public authority and individual rights and freedoms. Asian countries' constitutions seem to have similar main themes. As seen, topic 2 alludes to almost the same words as topic 2 of the European countries' constitutions. Therefore, both European and Asian countries' constitutions are similar to one another with respect to the largest topics. It may be suggested that to them both public authority and citizens' rights and freedoms are equally important.

As for the Middle Eastern countries, the largest topic with 219 similar words has to do with law, parliament, and state. The second largest topic with 211 words deals with citizens' rights and freedoms. African nations' constitutions are similar to those of the Middle Eastern countries with respect to the largest topic. As seen, the largest topic with 260 terms seems to focus on the national government itself and the rule of law. It is a little bit surprising that among the first five topics, no topic mentions of citizens' rights and freedoms, suggesting that the government itself and the structure of the government hold higher significance for the African nations. Middle Eastern and African nations are similar to one another with respect to the content of the largest topics of their constitutions.

The US constitution is different with respect to the content of the largest topic of its constitution. Looking at the largest topic, it may be argued that the US constitution focuses more on the role of both president and vice president, the office they hold, and the importance of the House.

5 Discussions and conclusions

A substantial portion of data is stored as text or unstructured data such as comments made by customers, digital libraries, email messages, and tweets. These unstructured data can be converted into structured data so that useful insights and hidden patterns may be extracted.

In this study, we took a text mining approach to analyze the constitutions of a number of countries located in Europe, Asia, Africa, and the Middle East and made a comparison between their constitutions and the US constitution. The quantitative data collected helped to uncover trends and patterns in the constitutions. These quantitative data were then visualized using the visualization tools in SAS to better understand the story behind the texts, numbers, and results.

It should be noted that text mining can indeed help data scientists analyze vast textual corpora fairly quickly by

automating the knowledge discovery process. However, it may also lead to ambiguity and loss of perception, which ultimately may lead to misinterpretations of the results.

One particular question we tried to answer in this study was "what are the key components in the constitutions of the countries in five different geographic regions? We were able to examine the patterns and trends within the constitutions to gather meaningful information. Our evaluation uncovered similarities and differences in the constitutions.

As expected, such important terms as "state," "law," and "president" are among the most frequently appearing terms throughout the constitutions across five different regions. Browsing through the regions and the terms, a number of interesting observations were made. For instance, certain terms such as "united," "congress," house," and "power" are among the most frequently appearing terms in the US constitution. However, they are not as common in the constitutions of the countries in other four regions. Similarly, while the terms "council," "public," and republic" are among the most commonly appearing terms in the constitutions of the Middle Eastern countries, no other regions' constitutions mention them as frequently. In addition, European nations in their constitutions allude to terms "federal," "provision," and "matter" a lot more frequently than any other four regions. Finally, while the term "assembly" appears 2221 times in the constitutions of African countries, it is not among the most frequently cited terms in other four regions. These differences may be attributed to the fact that different countries have different government structures and processes.

In addition, the term "act" holds the highest significance for the US constitution. Similarly, the term "intergovernmental" bears the highest importance for the European constitutions. For the Asian nations, the term "self-governing" seems to be holding more significance. The Middle Eastern counties assign the highest weight to the term "benches," and finally, the term "pillar" assumes the highest weight for African nations.

The cluster analysis and the topics generated in this study reveal that while some constitutions are more focused on the structure and the process of government along with citizens' rights and freedoms (European and Asian constitutions), African and Middle Eastern nations seem to focus more on the national government itself and the rule of law as no topic mentions of citizens' rights and freedoms, suggesting that the government itself and the structure of the government hold higher significance for them. The US constitution portrays a slightly different pictures as it focuses more on the role of both president and vice president, the office they hold, and the importance of the House.

This study suffers a couple of limitations. Firstly, although some regions such as Europe and Africa have a lot more countries, we picked only twenty countries from each region. Running a similar text mining analysis with a larger sample size would probably yield a different set of common terms with different weights. Secondly, some of the African and Asian countries surveyed in this study are former colonies of the European powers. Therefore, they were probably influenced by their colonial powers that once colonized and ruled them when drafting and penning their constitutions. These issues might be addressed in a follow-up study.

A comparative analysis provided in this study may be used as a frame of reference to gain a better understanding of how various countries' constitutions define and guarantee the fundamental rights of their citizens, and how they define the relationships between the branches of government. In other words, since this study reaches beyond one country and any specific culture, it provides a framework and context for the analysis of the differences and similarities between the constitutions of different nations located in different geographic regions. Finally, by running a comparative analysis of the constitutions of different countries by use of an empirical methodological framework, one can have a deeper understanding of how their own understanding of constitutional rights and responsibilities fit into the world around them.

Funding Not applicable.

Availability of data and materials The data used in this study are available on https://www.constituteproject.org/constitutions?lang=en.

Code availability Not applicable.

Declarations

Conflict of interest No conflict of interest.

References

Aggarwal CC, Zhai C (2012) An introduction to text mining. In: Aggarwal C, Zhai C (eds) Mining text data. Springer, Boston

Akilan A (2015) Text mining: Challenges and future directions. In: 2nd international conference on electronics and communication systems (ICECS), pp 1679–1684

Al-Daihani SM, Abraham AS (2018) Analysis of Academic Libraries' Facebook Posts: Text and Data Analytics. J Acad Librariansh 44(2):216–225

Andrews NO, Fox AE (2007) Recent developments in document clustering. http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf. Accessed 8 May 2018

Bigorra AM, Isaksson O, Karlberg M (2020) Semi-autonomous methodology to validate and update customer needs database through text data analytics. Int J Inf Manag 52:102073

Belokurows R (n.d) Comparing Brazilian and Portuguese constitutions with text mining. https://towardsdatascience.com/comparing-brazilian-and-portuguese-constitutions-with-text-mining-82213c7a95fd. Accessed 17 Jan 2021

Boudreaux DJ, Pritchard AC (1993) Rewriting the constitution: an economic analysis of the constitutional amendment process. Fordham Law Rev 62:111–162

Bulmer E (2017) What is a constitution? Principles and concepts. https://www.idea.int/sites/default/files/publications/what-is-a-constitution-primer.pdf. Accessed 11 Jan 2021

Britannica (2021) The editors of encyclopedia. Constitution of the United States of America. Encyclopedia Britannica, 2 Mar 2021. https://www.britannica.com/topic/Constitution-of-the-United-States-of-America. Accessed 31 May 2021.

Chakraborty G, Pagolu M, Satish G (2013) Text mining and analysis practical methods, examples and case studies using SAS. SAS Institute, Cary

Chen S, Guo X, Wu T, Ju X (2020) Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis. Inf Process Manag 57(5):102253

Chatterjee S, Goyal D, Prakash A, Sharma J (2021) Exploring healthcare/health-product ecommerce satisfaction: a text mining and machine learning application. J Bus Res 131:815–825

Galloway RW (1988) Basic constitutional analysis. Santa Clara Law Review 28(4):775–793

Gruss R, Abrahams AS, Ana W, Wang GA (2018) By the numbers: the magic of numerical intelligence in text analytic systems. Decis Support Syst 113:86–98

Hwang YI, Zheng L, Karystianis G, Gibbs V, Sharp K, Butler T (2020) Domestic violence events involving autism: a text mining study of police records in New South Wales, 2005–2016. Res Autism Spectr Disord 78:101634

IBM (2020) Text mining. https://www.ibm.com/cloud/learn/text-mining. Accessed 17 Jan 2021

Ibrahima NF, Wang X (2019) A text analytics approach for online retailing service improvement: evidence from Twitter. Decis Support Syst 121:37–50

Ittoo A, Nguyen ML, den Bosch A (2016) Text Analytics in industry: challenges, desiderata and trends. Comput Ind 78:96–107

Kang T, Park D, Han I (2018) Beyond the numbers: the effect of 10-K tone on firms' performance predictions using text analytics. Telematics Inform 35(2):370–381

Killian JH, Costello GA (1992) The Constitution of The United States of America Analysis and Interpretation, Retrieved January 17, 2021 from https://www.govinfo.gov/content/pkg/ GPO-CONAN-1992/pdf/GPO-CONAN-1992.pdf

Kobayashi VB, Mol ST, Berkers HA, Kismihok G, Den Hartog DN (2018) Text mining in organizational research. Organ Res Methods 21(3):733–765

Kwayu KM, Kwigizile V, Lee K, Oh J (2021) Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology. Accident Anal Prev 150:105899

Manning CD, Raghavan P, Schutze H (2008) Introduction to information retrieval, 1st edn. Cambridge University Press, New York

Pritchard AC, Zywicki TJ (1998) Finding the constitution: an economic analysis of tradition's role in constitutional interpretation. North Carolina Law Review 77:409–521

Reddick CG, Chatfield AC, Ojo A (2017) A social media text analytics framework for double-loop learning for citizen-centric public services: a case study of a local government Facebook use. Gov Inf Q 34(1):110–125

SAS (2021) SAS text miner reference help. SAS Institute Inc, Cary, NC, USA

Sohrabi B, Vanani RI, Narges N, Armin GR (2020) A predictive model of tourist destinations based on tourists' comments and interests using text analytics. Tourism Manag Perspectives 35:100710

Torii M, Fan JW, Yang Wl, Lee T, Wiley MT, Zisook DS, Huang Y (2015) Risk factor detection for heart disease by applying text analytics in electronic medical records. Journal of Biomedical informatics 58:S164–S170

Wand Y, Xu W (2018) Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decis Support Syst 105:87–95

Wolf W (2014) Text-mining South American constitutions. http://willwolf.io/2014/05/06/text-mining-south-american-constitutions/. Accessed 17 Jan 2021

Xiang Z, Schwartz Z, Gerdes JH, Uysal M (2015) What can big data and text analytics tell us about hotel guest experience and satisfaction? Int J Hosp Manag 44:120–130

Xu N, Ma L, Liu Q, Wang L, Deng Y (2021) An improved text mining approach to extract safety risk factors from construction accident reports. Saf Sci 138:105216

Zhang T, Bhatia A, Pandya D, Sahinidis NV, Cao Y, Flores-Cerrillo J (2020) Industrial text analytics for reliability with derivative-free optimization. Computers Chem Eng 135:106763