JISIB

Journal of Intelligence Studies in Business Vol. 13 No. 2 (2023) pp. 61-74
Open Access: Freely available at: http://jisib.com

Stevan Didijer
1911-2004

A comparative analysis with machine learning of public data governance and AI policies in the European Union, United States, and China

Christophe Bisson[*]
Skema France
christophe.bisson@skema.edu

Adele Giron
Skema France
adele.giron@skema.edu

Gauthier Verin
Skema France
gauthier1.verin@skema.edu

Received 24 July 2023 Accepted 12 September 2023

ABSTRACT This paper explores the public data governance and AI policies in the world's three main technological regions which are the United States, China, and European Union based on scientific literature analysis with machine learning. We used the RapidMiner text mining algorithm to classify texts and define the recuring themes in each region through Terms Frequency-Inverse Document Frequency, supervised machine learning techniques with KNN, and Naïve Bayes. Therein, our results reveal the most influential items for each region that emphasize three different approaches in China, the United States and the EU.

KEYWORDS: public data governance, artificial intelligence policy, text mining.

1. INTRODUCTION

Data has been construed as a key resource for companies and states pushing to be competitive (Kshetri, 2014). Furthermore, Mazurek and Malagocka (2019, 344) argue that "data may be seen as a currency in the digital world, and even compared to oil, gold or nowadays to labor." However, unlike other resources such as oil, data can be reused endlessly for different purposes and with unrestricted cross-border flows (Aaronson, 2019). Yet, data is often closely related to the notions of privacy and ethics, as it deals with human activities (Bisson, 2013).

The rise of the digital age was initially accompanied by the ambition to break down frontiers and create a "global village" (McLuhan, 1967). However, the digital transformation of many societies and the digitalization of human activities are challenging the concept of sovereignty. Yet, the conflict in Ukraine has further divided the world and ended the "fruitful" globalization that increased the importance of international public data governance policies. Therein, data storage and data infrastructures now crystallize the new spots of international geopolitical tensions and rivalries.

Various international regulations regarding data governance are also becoming weapons of economic warfare, as shown by

[*] *Corresponding author*

the China-U.S. rivalry (Zeng, 2020). Nowadays, to ensure the security and sovereignty of their data, states are adapting their laws to the digital sphere with technical and legal arsenals. Thus, data, AI technologies, and infrastructure networks today represent "a stake of sovereignty and power for states, international firms and other non-state actors" (Seurre, 2020, 3). Hence, digital technology is reshuffling the cards of the international power game and "data governance is becoming a political issue of crucial importance" (Matthews, 2019, 1) that can generate geopolitical conflicts. Furthermore, according to Zheng (2021, 1), policymakers must consider the concept of trilemma, i.e., "personal data protection, free transborder flow of information and the expansion of national jurisdiction" to build up new data transfer regulations.

In addition, the amount of data that companies have is essential for their development and contributes to sustaining their competitive advantages. Indeed, data that feeds their AI algorithms enables them to improve the efficiency and quality of their products and services, but it also helps them to create, or take part in the creation of, new ones.

According to the Data Governance Institute (2017), data governance is "a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods." Thus, data governance must be thought of as an ecosystem that encompasses privacy, security, ownership, use, and reuse of data, but also as the values and interests it contains (Winter and Davidson, 2017).

However, we cannot apply the same model of data governance in the private sector and in the public sector, as doing so requires political discussions (Okuyucu and Yavuz, 2020). As data governance involves various and numerous stakeholders, it is also closely linked to the concepts of sovereignty and data politics (Mureddu, Schmeling, and Kanellou, 2020). Woods (2018, 360) defines data sovereignty as the combination of "supreme control, over a territory, independent from other sovereigns." Yet, Liu (2021, 46) highlights the strategic role that data policies play in the "interactions between sovereign states or between the state and non-state actors over the collection, processing, transfer, sale, or use of data." Moreover, data policy is part of a much broader set of strategic digital policies, including the development of new technologies such as AI and its algorithms, and the security of their networks.

Data governance is closely related to privacy, as "data collection, analysis and processing are mainly perceived as a threat to privacy" (Mazurek and Malagocka, 2019, 349). Yet, Kuziemski and Misuraca (2020, 2) stress that "govern[ing] algorithms, while governing by algorithms" defines the very ambiguous situation that governments face, i.e., protecting their citizens by respecting their data privacy and security, and granting access to improve the efficiency of technological systems. Thus, designers of public data governance policies must consider individual privacy and data security, but also guarantee access to this data to improve the technology and keep innovating (Rosenbaum, 2010).

There is room for improvement regarding studies that deal with public data governance policies (Okuyucu and Yavuz, 2020; Liu, 2021). Yet, Alhassan, Sammon and Daly, (2017) highlight the need to investigate public data governance and AI policies. Moreover, Gleeson and Walden (2016) stress the public sector's dearth of maturity in this area.

Therein, to address this gap, we have used machine learning through RapidMiner text mining analysis to highlight the different approaches of public data governance and AI policies in the world's three main technological regions—China, the European Union, and the United States—based on scientific literature analysis. Moreover, by using machine learning to analyze most influential items related to public data governance and AI policies in each of the three regions, our research is congruent with Klyton et al. (2023, 142) that posit that "the language used by various stakeholders […] contributes to the construction of hegemonic power affecting (or supporting) organizational control".

In the remainder of this paper, we first highlight the relationship that exists between AI and data governance and the public data governance in the EU, United States, and China. Then we present our research design, discuss our results, and finish with our conclusion.

LITERATURE REVIEW

Artificial Intelligence as a tool for data governance

Artificial intelligence (AI) and data governance are interdependent and complementary (Matthews, 2019). The amount of data is fundamental when evaluating the accuracy of algorithms, whereas data quality and governance are essential prerequisites for the effective use of AI (Smith, 2019). The development of AI not only allows the transformation of data into action, but AI also learns from this information and creates new ones (Calzada and Almirall, 2020). Yet, AI relies above all on the strategies and public policies that governments put in place using their companies as application tools.

The development of AI and other machine-learning systems brings new challenges for data governance: "the scale and scope of data used by the algorithms and the opacity of the algorithms" (Winter and Davidson, 2017, 281) regarding the way data is used and transformed into results. AI has "the increased capability to collect, analyze and combine vast amounts of data from different sources, […] thus enhancing the capabilities of technology powers" (Mazurek and Malagocka, 2019, 348). AI can work on specific tasks without human monitoring, which enhances its analysis performance. Therefore, the definition of AI directly implies data privacy issues, as AI can easily deduce or predict sensitive and personal information.

However, data governance policies are far more difficult to implement nowadays because data can be collected through numerous different devices (e.g., smartphones, watches, GPS) that do not belong to a state but to a private company that can sell the data to another country. In this context, it is not the public authorities that own real-time data, but private companies such as network operators or big technological firms (Alemanno, 2018). Thus, the enormous amount of data and its stakeholders increase the difficulty of implementing a strict and efficient data governance policy. Therefore, international data sharing is a source of economic and commercial development, but it also represents a risk to the privacy of citizens who are potentially vulnerable to "foreign surveillance, hacking and data breaches" (Liu, 2021, 51).

EU, U.S., and Chinese public data governance: three different approaches

The General Data Protection Regulation: a strong but inadequate European regulation

The fast-moving digital environment and exponential growth of data has led the EU to implement strong regulations to protect its citizens' privacy: the General Data Protection Regulation (GDPR). Yet, very recently the EU parliament adopted the 'AI Act' that will be voted now in the EU council aiming to reach an agreement by the end of 2023 to become a law (EU Parliament, 2023).

In Europe, the protection of citizens' personal data constitutes a fundamental right because it is inherently related to a natural person. In doing so, data protection means: "preserving a natural person from the misuse of his or her personal information" (Fabiano, 2019, 58).

Implemented in 2018, the GDPR regulation demands that companies comply with a set of rules regarding the collection, storage, and processing of European's personal data in response to ethics and privacy concerns. These measures imply for companies major technological, functional, juridical, and cultural changes and challenges, not without impacts on businesses. These far-reaching changes are time-consuming and expensive—necessitating the mobilization of additional financial and human resources—and they represent high barriers for companies which must adapt their overall organization (processes, routines, procedures) to ensure that their activities meet GDPR rules.

Thereby, to ensure data protection in cross-border flows, GDPR regulation has been added in every international trade agreement signed with the EU. Instead of applying different regulations according to the trading country, companies implement the European data protection law on a global scale. Therefore, the GDPR gets an extraterritorial scope as the foreign market follows this regulation with their partner countries in almost all their data transmission flows.

In practice, this regulation provides European citizens with the right to obtain an explanation regarding the decisions made by algorithms. However, Gordon-Murnane (2018, 41) highlights that "the GDPR lacks precise language as well as explicit and well-defined rights and safeguards against automated decision-making, and therefore runs the risk of being toothless." Hence, in the long run, the GDPR does not guarantee policies and technologies that comply with ethics and privacy. To be efficient, data protection law must be implemented in every step of the new technologies' development, from early stages to final processing. Unfortunately, our policymakers, industries, academia, and public sectors do not have the necessary resources to both develop efficient and innovative technologies and respect data protection law (Panagiotopoulos, 2019).

Yet, European regulators struggle to catch up and keep up with the growing pace of our digital evolution, and many measures that the GDPR has implemented conflict with the way global computer networks currently operate. The right to erase personal data is one of the most complex conditions to meet (Teixeira, Mira da Silva and Pereira 2019). In his report, Herian (2020) highlights this paradox, i.e., the right for European citizens to erase their personal data and the prohibition of storing personal data, which negatively impacts the possibility of improving control over one's own personal data.

Under GDPR regulation, data processing is governed by the purpose limitation concept: Data is collected and used only for a clearly defined purpose. However, the core principle of big data is to collect and analyze a very large amount of data and use it for different purposes, rather than a single defined purpose (Okuyucu and Yavuz, 2020). Thus, even if the data is not personal, platforms (e.g., Facebook) could reidentify a person by analyzing and triangulating enough data (Bendiek and Römer, 2018). As AI relies on big data systems, GDPR regulation on this technology is often discussed, with critics arguing both that this new law will slow down and limit innovation and that it threatens individual freedoms and fundamental human rights.

The Clarifying Lawful Overseas Use of Data Act: an instrument of U.S. digital supremacy

In democratic and liberal societies, privacy and data protection are tied to liberty. However, in the United States, privacy and data protection regulations are seen in another light, as "the focus is not on the protection of human dignity, but on freedom in the sense of liberty as a civil right of the individual, who wishes to be free of legal regulations" (Bendiek and Römer, 2018, 37). These different concepts can lead to conflicts of interest between states.

Implemented in 2018, the Clarifying Lawful Overseas Use of Data Act (CLOUD Act) enables the U.S. government to access every data center owned by a U.S. company, regardless of the geolocation of the servers. The CLOUD Act gives U.S. authorities permission to access U.S. cloud providers and request data on U.S. or foreign citizens and companies without their prior consent (Brincourt, 2021). Through bilateral agreements with national governments, the United States may also have access to data stored in the partner country. Thus, the CLOUD Act and the GDPR both have an extraterritorial characteristic involving numerous conflicts over the regulation of governance and digital sovereignty (Bendiek and Römer, 2018).

The CLOUD Act was enacted "in the name of protecting the public safety of the United States and fighting the most serious infractions, crimes, and terrorism" (U.S. Department of Justice, 2021). But this extraterritorial federal law expressly reflects the willingness of U.S. authorities to have access to data stored at service providers (Duboc and Noël, 2021). In accordance with a simple judge approval, service providers must communicate "the contents of electronic communications, any record, any information relating to a customer or subscriber, including personal data. The person who owns the data will not be notified" (The U.S. Department of Justice, 2021). The general characteristics of the law grant it a very broad scope of application:

individuals, companies, and the state, on all exchanges or data, wherever they are stored. This law breaches regulations concerning personal data protection, corporate data protection, and the protection of highly confidential elements of strategic state security. Considering that 85% of the global digital storage market is operated by U.S. firms, almost all European companies are potentially affected by this policy (Duboc and Noël, 2021).

China's data governance policies: a pillar of the Middle Kingdom's digital power

Liu (2021, 48) posits, "the Chinese government has long realized the strategic value of data." Since Xi Jinping came to power in 2013, the Chinese government has been building its big data national strategy and forging its way to becoming a "cyber superpower" (Segal, 2017) through:

- Strategic collaboration with digital firms
- Integration of big data into government statistics
- Upgrading of big data to a national strategy level
- Construction of a massive national data center

In 2017, the Chinese president enacted the Intelligence Law, its intention being to "strengthen the ability to protect the nation's crucial data resources, speed up relevant legislation, and improve protection of data property rights." (China Daily Newspaper, 2017). The extremely blurred framework of this law may give rise to fears of its extensive application, as with the U.S. CLOUD Act (Duboc and Noël, 2021).

China is continuously updating its data governance strategy and agenda, which makes it more relevant to its external environment. In China, this concept is intrinsically linked to the broader notion of "cyber sovereignty," defined by Liu (2021, 52) as "state control of digital technologies, content and infrastructure under their jurisdiction." The six fundamental texts of the Chinese data strategy are: the cybersecurity law (2016), the information security technology guidelines for data cross-border transfer security assessment (2017), the draft measures on security assessment of the cross-border transfer of personal information (2019), the draft measures for data security management (2019), the data security law of the PRC (2021), and the personal information protection law of the PRC (2021).

Early in the summer of 2022, China enhanced its data governance policy, and especially the data cross-border arrangement. Since then, companies have had to follow a set of CAC-implemented rules to transfer abroad any data created in PRC. Any company operating in the country must store its data in China. If a data transfer is to be made to another country, the government conducts a preliminary risk assessment. China's data localization requirement applies to the personal data of Chinese nationals, but also to so-called "important data," a blurred definition that can include any type of data, and that is all available to the Chinese government.

Hence, the approaches in this matter as are embodied in the EU, U.S., and Chinese regulations result "in three inherently incompatible legislative paradigms, which has led to the restricted flow of personal data around the world as well as the free flow in three different regions, with the EU, the United States and China as the center of each region" (Zheng, 2021, 1).

RESEARCH DESIGN

Sampling and document preprocessing

We initially made requests on a selection (to which we had full access) of online databases— "Emerald Insights," "Open Edition," and "Science Direct"—to gather scientific documents about "public data governance" and "AI policy(ies)" and "EU" or "China" or "United States." This allowed us to get 62 documents as a sample. We used RapidMiner text mining software to conduct our text analysis of these 62 documents. For that purpose, we needed to correctly prepare documents to get the appropriate formats, clean the database, and obtain more meaningful results. Yet, we performed a tokenization of our texts by "non-letters" to obtain non-letter characters as segment cuts in our texts. As we changed the texts from PDF format to TXT format, the non-letter characters could only be spacing or dashing, therefore segmenting our texts by words. Thereafter, we undertook a data "cleaning" in our database before running the analysis.

First, we filtered the tokens by length, (4 to 20 characters) and by the stop words. Finally, we performed the stemming (snowball) process to restore all the words back to their roots (e.g., *connect*, *connections*, and *connected* all include the root *connect*).

Moreover, before processing the Terms Frequency-Inverse Document Frequency (TF-IDF) analysis, we labeled our texts now turned into segments by using their origins i.e., China, EU, and United States.

Methods

Terms Frequency-Inverse Document Frequency (TF-IDF)

Terms Frequency-Inverse Document Frequency (TF-IDF) is a common method used in text mining to retrieve information (Christian et al., 2016). Its goal is to show the importance of each word to a document in a corpus. Compared to a basic word count, TF-IDF helps to underline the most important word in each label. In this case, we used the prune method to take only into account terms that appear between 5 to 9,999 times in the corpus.

KNN algorithm

The K-Nearest Neighbor (KNN) algorithm is a supervised machine-learning algorithm mainly used for classification (Uddin et al., 2022). Our aim was to utilize KNN to automatically classify the documents by nationality with this content and see how accurate it was. As we worked on a small amount of data, we wanted the data set that we got from the TF-IDF operations to be at the same time the training set and the test set. To do so, we used cross validation, i.e., in our case we used 80% of the data set to train the algorithm and 20% to test the result accuracy. We defined a stratified sampling to determine which texts would be used as part of the training set and which would be used as part of the test set, while having the same percentage of each subset (here labeled as the country) in the training and in the test set. Yet, we utilized the cosine similarity between two documents based on their TF-IDF score, which we calculated previously to determine the similarities of the texts. Then, we determined the right number of K Neighbor.

Naïve Bayes algorithm

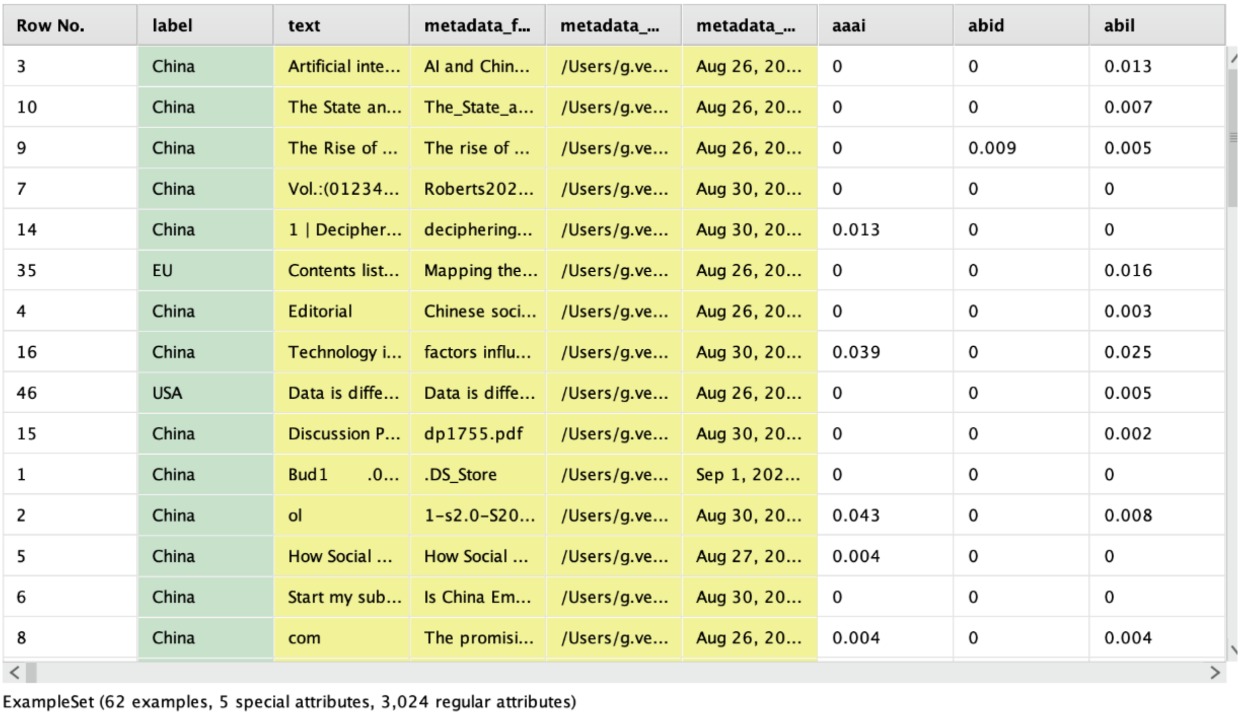
The Naïve Bayes algorithm allowed us first to come up with another system of classification to perform another accuracy that might do better than the KNN one (Prasad et al., 2022). Thus, we could investigate which word(s) the algorithm found more determining for each label to classify them as a Chinese, American, or European texts. Yet, we did the Laplace correction, since if an event never occurs then its conditional probability would be equal to 0; therefore, we added 1 to each count feed in the algorithm (Wang et al., 2022).

RESULTS AND DISCUSSION

TF-IDF results

When we ran the TF-IDF algorithm, it resulted in a table containing 62 rows (one for each document), and 3,029 columns, column 2 being the labels, columns 4, 5, 6 containing the metadata and the other 3,024 containing single terms with the TF-IDF scores. The scores are low as we have many terms in each document (see Table 1).

Table 1. TF-IDF results

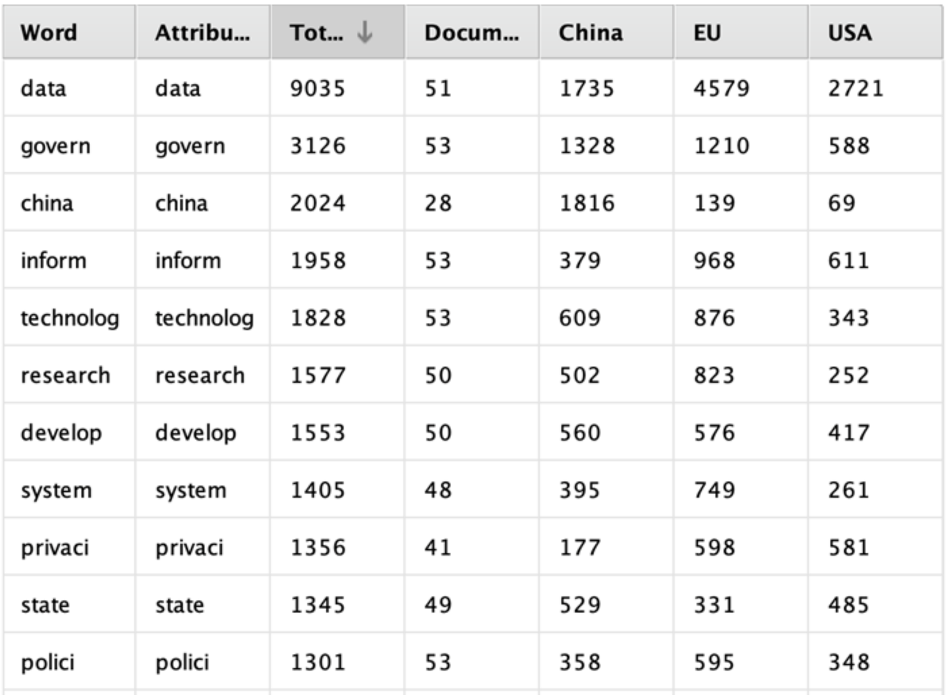


Row No.	label	text	metadata_f...	metadata_...	metadata_...	aaai	abid	abil
3	China	Artificial inte...	AI and Chin...	/Users/g.ve...	Aug 26, 20...	0	0	0.013
10	China	The State an...	The_State_a...	/Users/g.ve...	Aug 26, 20...	0	0	0.007
9	China	The Rise of ...	The rise of ...	/Users/g.ve...	Aug 26, 20...	0	0.009	0.005
7	China	Vol.:(01234...	Roberts202...	/Users/g.ve...	Aug 30, 20...	0	0	0
14	China	1 | Decipher...	deciphering...	/Users/g.ve...	Aug 30, 20...	0.013	0	0
35	EU	Contents list...	Mapping the...	/Users/g.ve...	Aug 26, 20...	0	0	0.016
4	China	Editorial	Chinese soci...	/Users/g.ve...	Aug 26, 20...	0	0	0.003
16	China	Technology i...	factors influ...	/Users/g.ve...	Aug 30, 20...	0.039	0	0.025
46	USA	Data is diffe...	Data is diffe...	/Users/g.ve...	Aug 26, 20...	0	0	0.005
15	China	Discussion P...	dp1755.pdf	/Users/g.ve...	Aug 30, 20...	0	0	0.002
1	China	Bud1 .0...	.DS_Store	/Users/g.ve...	Sep 1, 202...	0	0	0
2	China	ol	1-s2.0-S20...	/Users/g.ve...	Aug 30, 20...	0.043	0	0.008
5	China	How Social ...	How Social ...	/Users/g.ve...	Aug 27, 20...	0.004	0	0
6	China	Start my sub...	Is China Em...	/Users/g.ve...	Aug 30, 20...	0	0	0
8	China	com	The promisi...	/Users/g.ve...	Aug 26, 20...	0.004	0	0.004

ExampleSet (62 examples, 5 special attributes, 3,024 regular attributes)

Table 2 indicates the 11 most frequently appearing words in our documents, allowing one to see the main subjects of our research, i.e., data governance, research, privacy, state, and policies. All those terms are important as they are the main subjects of most papers. However, it would be difficult to use this data for a classifier algorithm as it would use those terms as important terms, even though they are not discriminant.

Table 2. Most frequently appearing words



Word	Attribu...	Tot... ↓	Docum...	China	EU	USA
data	data	9035	51	1735	4579	2721
govern	govern	3126	53	1328	1210	588
china	china	2024	28	1816	139	69
inform	inform	1958	53	379	968	611
technolog	technolog	1828	53	609	876	343
research	research	1577	50	502	823	252
develop	develop	1553	50	560	576	417
system	system	1405	48	395	749	261
privaci	privaci	1356	41	177	598	581
state	state	1345	49	529	331	485
polici	polici	1301	53	358	595	348

We then fed those results into the KNN classifier.

KNN classifier results

First, we needed to determine the optimum parameter number of K neighbor. As we built a loop to be more efficient, we could thereby determine which number of K neighbor has the minimum error rate. Our results stress that the fourth iteration has the minimum error rate: 27%. Therefore, we know that 5 K neighbor is the optimum parameter with an accuracy of 73% (see Figure 1).

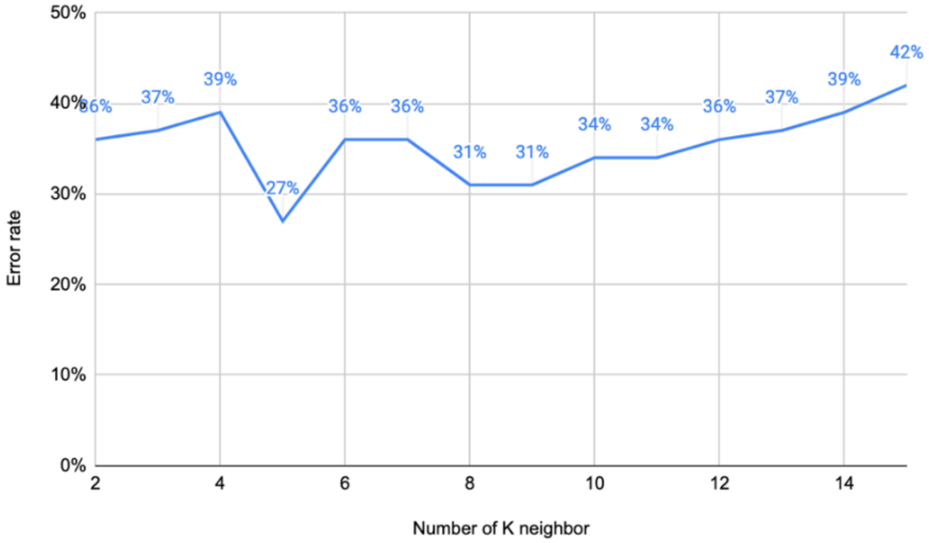
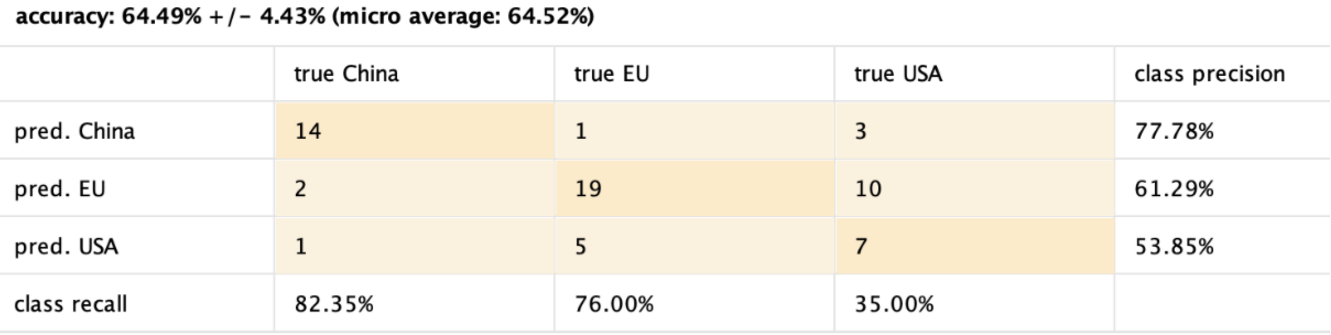


Figure 1. Error rate number of K neighbor

Thereafter, the confusion matrix of our KNN algorithm with 5 K neighbor set as parameter was done. As Table 3 indicates, the algorithm performs very well with predicting text coming from China—a 94.12% accuracy, with only one error of a European text predicted as Chinese. Hence, Chinese texts are highly recognizable, as they have a clear tendency on the topics (half of texts deals with Covid-19 and AI). Most of the Chinese texts mention the role of the state and its tight influence on research and companies.

The KNN algorithm performs rather well with the European text, with an accuracy of 80%. However, the accuracy drops below 50% for the American texts, with a tendency to label similarly to European texts, predicting nine texts as European and nine correctly predicted as American. The fact that there are also more European texts than American might additionally create a negative effect for the classifier, which has less data to train with to recognize the American label.

Table 3. KNN confusion matrix

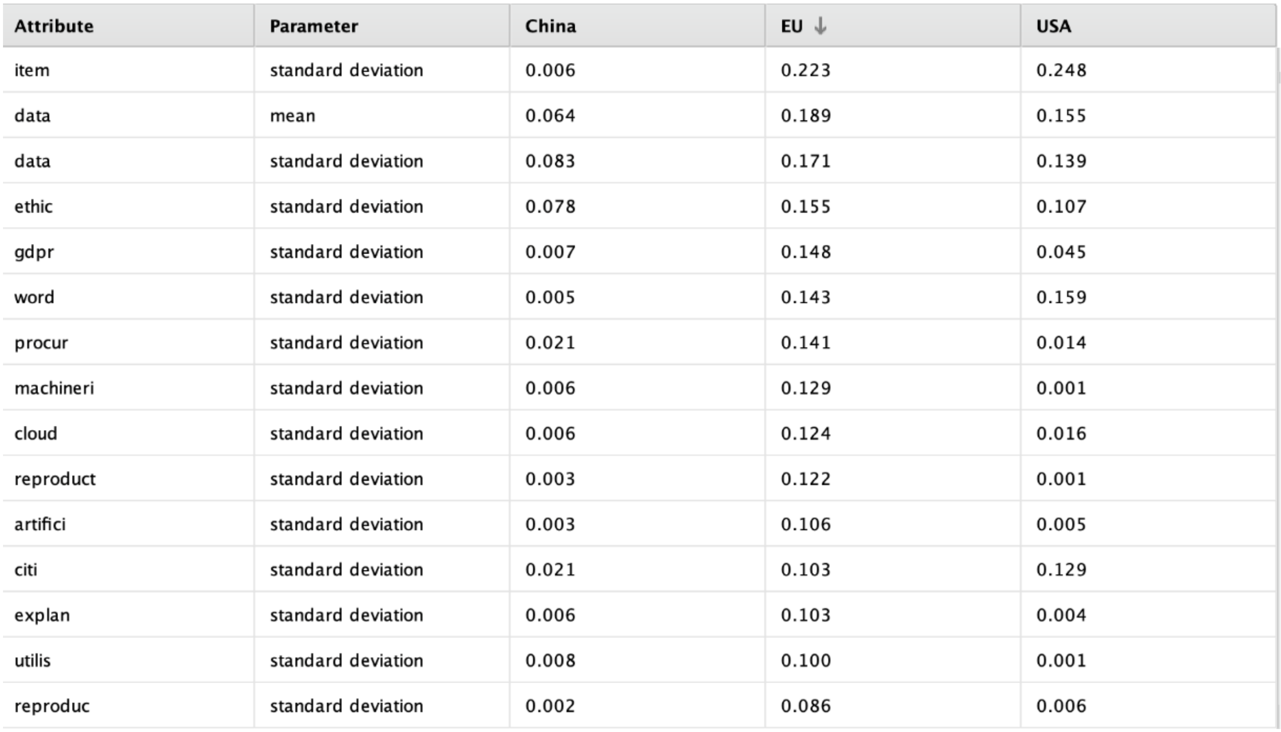


accuracy: 64.49% +/- 4.43% (micro average: 64.52%)

	true China	true EU	true USA	class precision
pred. China	14	1	3	77.78%
pred. EU	2	19	10	61.29%
pred. USA	1	5	7	53.85%
class recall	82.35%	76.00%	35.00%	

Naïve Bayes algorithm results

Table 4. Naïve Bayes confusion matrix

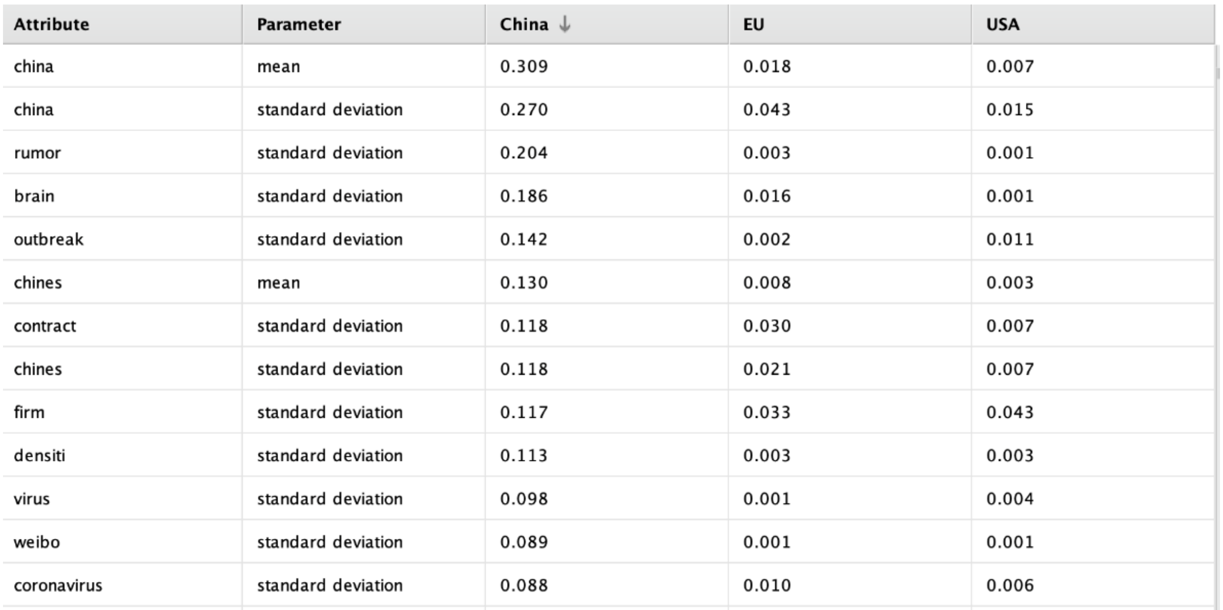


Attribute	Parameter	China	EU ↓	USA
item	standard deviation	0.006	0.223	0.248
data	mean	0.064	0.189	0.155
data	standard deviation	0.083	0.171	0.139
ethic	standard deviation	0.078	0.155	0.107
gdpr	standard deviation	0.007	0.148	0.045
word	standard deviation	0.005	0.143	0.159
procur	standard deviation	0.021	0.141	0.014
machineri	standard deviation	0.006	0.129	0.001
cloud	standard deviation	0.006	0.124	0.016
reproduct	standard deviation	0.003	0.122	0.001
artifici	standard deviation	0.003	0.106	0.005
citi	standard deviation	0.021	0.103	0.129
explan	standard deviation	0.006	0.103	0.004
utilis	standard deviation	0.008	0.100	0.001
reproduc	standard deviation	0.002	0.086	0.006

The confusion matrix of the Naïve Bayes' classifier (see Table 4) didn't perform as well as the KNN classifier: 64.5% performance (as average). The algorithm performs less well when classifying all three labels, with a 12% drop in classification of the Chinese ones.

To investigate which terms influence each label the most, we checked the distribution table. This table shows two values per attribute, i.e., the mean deviation and the standard deviation. Our results highlight that there are more similarities among the influential terms between Europe and the United States (see Table 5), and the most influential word for the U.S. label also has strong influence on the EU one (thereby the classifier struggles to differentiate between the two classes).

Table 5. The EU sort Naïve Bayes distribution table

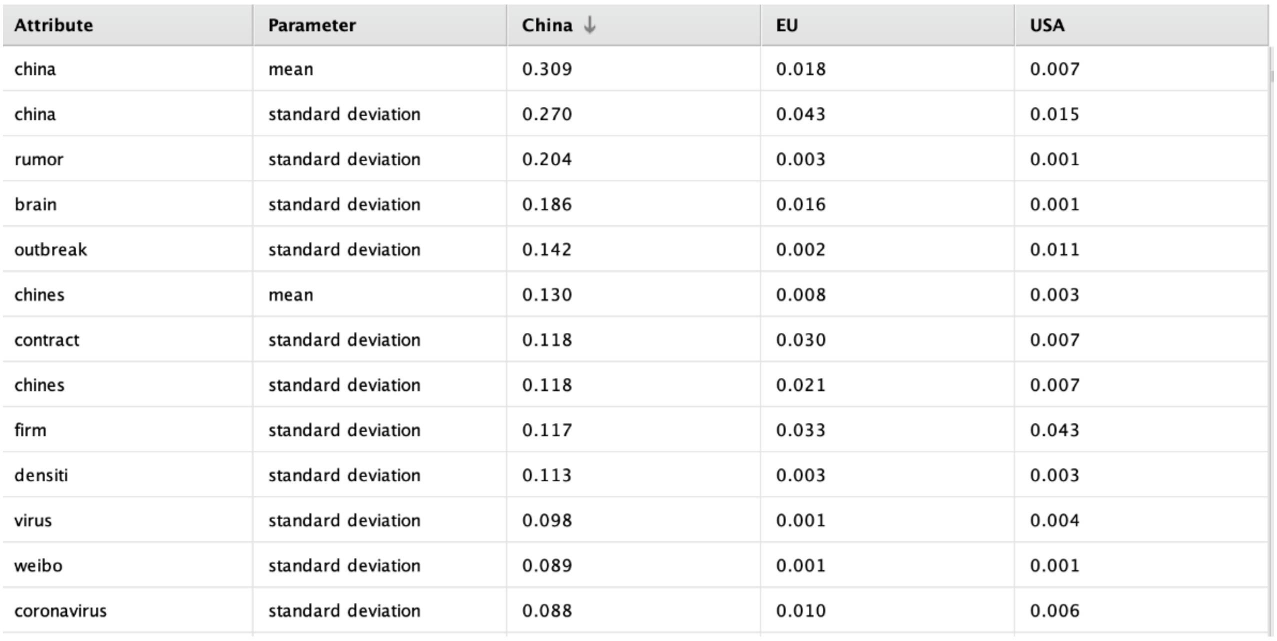


Attribute	Parameter	China ↓	EU	USA
china	mean	0.309	0.018	0.007
china	standard deviation	0.270	0.043	0.015
rumor	standard deviation	0.204	0.003	0.001
brain	standard deviation	0.186	0.016	0.001
outbreak	standard deviation	0.142	0.002	0.011
chines	mean	0.130	0.008	0.003
contract	standard deviation	0.118	0.030	0.007
chines	standard deviation	0.118	0.021	0.007
firm	standard deviation	0.117	0.033	0.043
densiti	standard deviation	0.113	0.003	0.003
virus	standard deviation	0.098	0.001	0.004
weibo	standard deviation	0.089	0.001	0.001
coronavirus	standard deviation	0.088	0.010	0.006

However, the most-influential Chinese terms are unique to its label. With almost no influence on the two other classes, this explains why China is the best-labeled category, with an error rate of 18% (see Table 6). It is congruent with Zeng (2020, 1442) who highlighted “the successful employment of digital technologies in China is made possible by [China’s] unique socio-political environment.” Thus, for China, we observe many terms related to the Covid-19 pandemic (e.g., *outbreak*, *virus*, *coronavirus*, *epidemic*). The term *authoritarian* appears in Row 16, thereby allowing them to focus only on the technical part of AI, and not on public opinion nor law limitations, which is in line with what was stated in our literature (Zeng, 2020; Liu, 2021).

Table 6. The Chinese sort Naïve Bayes distribution table

If we add up the 15 most influential attributes for China, we reach the number 2.155; the same attributes for the EU reach 0.205 and 0.160 for the United States. This means that the influential terms for China are unique for this specific label. If we perform the same analysis on the EU, we find that its 15 most influential terms reach the number 2.041, which is quite a shift when compared with China, for which those terms score 0.321. However, those terms do have an influence on the U.S. label, where they score 1.029. This confirms that the difference between the EU and U.S. texts is not as wide as the difference between the EU and the Chinese texts. Moreover, the 15 most influential terms for the United States and EU share three terms: *data*, *item*, *word*.



Attribute	Parameter	China ↓	EU	USA
china	mean	0.309	0.018	0.007
china	standard deviation	0.270	0.043	0.015
rumor	standard deviation	0.204	0.003	0.001
brain	standard deviation	0.186	0.016	0.001
outbreak	standard deviation	0.142	0.002	0.011
chines	mean	0.130	0.008	0.003
contract	standard deviation	0.118	0.030	0.007
chines	standard deviation	0.118	0.021	0.007
firm	standard deviation	0.117	0.033	0.043
densiti	standard deviation	0.113	0.003	0.003
virus	standard deviation	0.098	0.001	0.004
weibo	standard deviation	0.089	0.001	0.001
coronavirus	standard deviation	0.088	0.010	0.006

Yet, it is very interesting to find words such as *GDPR* and *ethic* as the fourth and fifth most influential terms for European-labeled texts. As defined in the literature review, there is a difference between the way China and EU countries are working on AI. The EU is trying to build a different model based on the spectrum of data privacy (first and thirty-first terms ranked) and prioritize the rights of its citizenry (city ranked 12th). This, according to Hlávka (2020), is one of the reasons why the EU lags in terms of AI-related technological advances. Furthermore, this is where we can see the difference between the EU and the United States.

Even though it is not as relevant as the differences that the EU and the United States have with China, there is still a difference between the United States and the EU. Our algorithms allow us to obtain more terms coming from U.S. private industry, as well as potential application of AI (same as China, and a few words being related to Covid-19), as we have medical terms like *health* ranked fourth, *healthcare* ranked seventh, *medic* 11th, and *patient* 13th—but also, terms that are related to company property, such as *court*, *venture*, and *copyright*. Therefore, as Duboc and Noel (2021) stressed, even if the United States created the CLOUD Act (which is still very vague on different subjects), it still allows massive collections of data—wherever their tech companies do business in the world— to feed their algorithm and maintain their competitiveness against the Chinese tech companies that are government-sponsored.

CONCLUSION

Data constitutes the main vector today of the success of companies and countries. Yet, data represents "the most important factor to ensure successful AI algorithms" (Lee, 2018, 34) as it "feeds" the AI algorithm constructed. AI importance is growing, as pointed out in 2017 by Russia's President Putin: "whoever leads in artificial intelligence will rule the world." (Meyer, 2017). Therein, to control the internet and data, states define public data governance policies (Woods, 2018). Alhassan et al. (2017) amplified the need to investigate public data governance and AI policies. In an aim to help address this gap, we've used machine learning through RapidMiner text mining analysis to highlight the different approaches to public data governance and AI policies as they exist in the Chinese, European, and U.S. literature. We've sought to determine whether a classification of the texts obtained in scientific databases is possible according to the key words characterizing the approach to public data governance and AI policies and depending on the three geographical areas selected.

We obtained a 72% accuracy with the KNN algorithm using the cosine similarity with the number of neighbors set to 5- and it performed well on the Chinese' texts and less well when differentiating U.S. labeled text from European text.

We used the Naïve Bayes algorithm and obtained a 64% accuracy, which is not as good as that which was achieved with KNN. However, it enabled us to understand better how the algorithm weighted each probability to classify the text. We determined that Chinese top discriminant terms were more unlikely to also be discriminant for the EU and U.S. texts. While the EU and U.S. texts tend to be more similar and so have similar discriminant terms.

Our results emphasize that China has no legal limit in terms of developing its big database's algorithm. Yet, the United States tends to focus more on its data sovereignty, but with more mentions of ethics or privacy than in China. Regarding the EU, it highlights that the EU is trying to build a model that focuses on data privacy and rules to protect its citizens' privacy.

The survey has some limitations due to the limited number of databases used as well as the limitations of RapidMiner.

REFERENCES

Aaronson, S.A. 2019. Data is different, and that's why the world needs a new approach to governing cross-border data flows. Digital Policy, Regulation and Governance 21(5): 441–460.

Alemanno, A. 2018. Big Data for Good: Unlocking Privately-Held Data to the Benefit of the Many. European Journal of Risk Regulation 9(2): 183–191.

Alhassan, I., Sammon, D. and Daly, M. 2018. Data governance activities: a comparison between scientific and practice-oriented literature. Journal of Enterprise Information Management 31(2): 300–316.

Bendiek, A. and Römer, M. 2019. Externalizing Europe: the global effects of European data protection. Digital Policy, Regulation and Governance 21(1): 32–43.

Bisson, C. 2013. Guide de gestion strategique de l'information pour les PME. Montmoreau : Les 2Encres.

Brincourt, L. 2021. Géopolitique de la Datasphère. Le "Cloud Act", trois ans après : révélateur du besoin de définition de notre souveraineté dans l'espace numérique. Accessed November 17, 2022. https://www.diploweb.com/Le-Cloud-Act-trois-ans-apres-revelateur-du-besoin-de-definition-de-notre-souverainete-dans-l-espace.html

Calzada, I. and Almirall, E. 2020. Data ecosystems for protecting European citizens' digital rights. Transforming Government: People, Process and Policy 14(2): 133–147.

China Daily Newspaper. 2017. 习近平：实施国家大数据战略加快建设数字中国. Accessed November 21, 2022. http://www.chinadaily.com.cn/interface/flipboard/1142846/2017-12-12/cd_35280418.html.

Christian, H., Pramodana Agus, M and Suhartono, D. 2016. Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF). ComTech: Computer, Mathematics and Engineering Applications 7(4): 285-294.

Data Governance Institute 2022. Definitions of Data Governance. Accessed November 21,2022. https://datagovernance.com/the-data-governance-basics/definitions-of-data-governance/

Duboc, S. and Noël, D.-J. 2021. Economie et gouvernance de la donnée. Accessed November 17, 2022. https://www.lecese.fr/sites/default/files/pdf/Avis/2021/2021_06_eco_gouv_donnee.pdf

EU Parliament. 2023. EU AI Act: first regulation on artificial intelligence. Accessed June 25, 2023. EU AI Act: first regulation on artificial intelligence | News | European Parliament (europa.eu)

Fabiano, N. 2019. Ethics and the Protection of Personal Data. Journal of Systemics Cybernetics and Informatics 17(2): 58–64.

Gleeson, N. and Walden, I. 2016. Placing the state in the cloud: Issues of data governance and public procurement. Computer Law & Security Review 32(5): 683–695.

Gordon-Murnane, L. 2018. Ethical, Explainable Artificial Intelligence - Bias and Principles. Online Searcher 42(2): 22–44.

Herian, R. 2018. Regulating Disruption: Blockchain, GDPR, and Questions of Data Sovereignty. Social Science Research Network 22(2): 7–16.

Hlávka, J.P. 2020. Security, privacy, and information-sharing aspects of healthcare artificial intelligence. In Artificial Intelligence in Healthcare, ed. A. Bohr and K. Memarzadeh, 235–270. Amsterdam, Netherlands: Elsevier.

van Klyton, A. Arrieta-Paredes, M.-P., Palladino, N. and Soomaree, A. 2023. Hegemonic practices in multistakeholder Internet governance: Participatory evangelism, quiet politics, and glorification of status quo at ICANN meetings. The Information Society 39 (3): 141-157.

Kostka, G. 2019. China's social credit systems and public opinion: Explaining high levels of approval. New Media & Society 21(7): 1565–1593.

Kshetri, N. 2014. Big data׳s impact on privacy, security and consumer welfare. Telecommunications Policy,38(11): 1134–1145.

Kuziemski, M. and Misuraca, G. 2020. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. Telecommunications Policy 44(6): 101976.

Lee, K. 2018. AI Superpowers: China, Silicon Valley, And The New World Order. Houghton: Mifflin Harcourt Company.

Liu, L. 2021. The Rise of Data Politics: Digital China and the World. Studies in Comparative International Development 56(1): 45–67.

Matthews, K. 2019. How AI is slowly changing data governance. Information Management 27 November.

Mazurek, G. and Małagocka, K. 2019. Perception of privacy and data protection in the context of the development of artificial intelligence. Journal of Management Analytics 6(4): 344–364. McLuhan, M., & (1967). The medium is the massage: An inventory of effects.

McLuhan, M. and Fiore, Q. 1967. The Medium is the Massage: An Inventory of Effects. Berkeley, CA: Gingko Press.

Meyer, D. 2017. Vladimir Putin Says Whoever Leads in Artificial Intelligence Will Rule the World. Accessed Oct 20, 2022. https://fortune.com/2017/09/04/ai-artificial-intelligence-putin-rule-world/,

Mureddu, F., Schmeling, J. and Kanellou, E. 2020. Research challenges for the use of big data in policy-making. Transforming Government: People, Process and Policy 14(4): 593–604.

Okuyucu, A. and Yavuz, N. 2020. Big data maturity models for the public sector: a review of state and organizational level models. Transforming Government: People, Process and Policy 14(4): 681–699.

Panagiotopoulos, A. 2019. Data protection law and ethics: Where do we stand? Information and records management society (212): 8–11.

Prasad, R., Agrawal, R. and Sharma, H. 2022. Modified Gabor Filter with Enhanced Naïve Bayes Algorithm for Facial Expression Recognition in Image Processing. In Advances in Computational Intelligence and Communication Technology, Lecture Notes in Networks and Systems, ed Gao, XZ., Tiwari, S., Trivedi, M.C., Singh, P.K., Mishra, K.K., 371–383. Singapore: Springer.

Rosenbaum, S. 2010. Data Governance and Stewardship: Designing Data Stewardship Entities and Advancing Data Access. Health Services Research 45(5): 1442–1455.

Segal, A. 2021. When China Rules the Web: Technology in Service of the State. Foreign Affairs 24 November.

Seurre, X. 2020. L'intelligence artificielle, un enjeu stratégique pour la puissance chinoise. Accessed November 17, 2022. https://www.iris-france.org/notes/lintelligenceartificielle-un-enjeu-strategique-pour-la-puissance-chinoise/

Smith, D. 2019. AI in Data Governance Strategic Finance, 1 November. Accessed November 17, 2022. https://sfmagazine.com/post-entry/september-2019-ai-in-data-governance/

Teixeira, A. G., Mira da Silva, M. and Pereira, R. 2019. The critical success factors of GDPR implementation: a systematic literature review. Digital Policy, Regulation and Governance 21(4): 402–418.

Uddin, S., Haque, I., Lu, H., Moni, M.A. and Gide, E. 2022. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep 12, 6256.

The US Department of Justice. 2021. Cloud act resources. Accessed November 22, 2022. https://www.justice.gov/dag/cloudact

Wang Y, Li T, Liu M, Li C, Wang H. 2022. Study on token shuffling under incomplete information based on machine learning. International Journal of Intelligent systems 37(12): 1-23.

Winter, J.S. and Davidson, E. 2018. Big data governance of personal health information and challenges to contextual integrity. The Information Society,35(1): 36–51.

Woods, A.K. 2018. Litigating Data Sovereignty. Yale Law Journal 128(2): 328–406.

Zeng, J. (2020).Artificial intelligence and China's authoritarian governance. International Affairs 96(6): 1441–1459.

Zheng, G. 2021. Trilemma and tripartition: The regulatory paradigms of cross-border personal data transfer in the EU, the U.S. and China. Computer Law &Amp; Security Review 43: 105610.