What do governments plan in the field of artificial intelligence? Analysing national AI strategies using NLP

Theodoros Papadopoulos
University of the Aegean
Greece
t.papadopoulos@aegean.gr

Yannis Charalabidis
University of the Aegean
Greece
yannisx@aegean.gr

ABSTRACT

The primary goal of this paper is to explore how Natural Language Processing techniques (NLP) can assist in reviewing, understanding, and drawing conclusions from text datasets. We explore NLP techniques for the analysis and the extraction of useful information from the text of twelve national strategies on artificial intelligence (AI). For this purpose, we are using a set of machine learning algorithms in order to (a) extract the most significant keywords and summarize each strategy document, (b) discover and assign topics to each document, and (c) cluster the strategies based on their pair-wise similarity. Using the results of the analysis, we discuss the findings and highlight critical issues that emerge from the national strategies for artificial intelligence, such as the importance of the data ecosystem for the development of AI, the increasing considerations about ethical and safety issues, as well as the growing ambition of many countries to lead in the AI race. Utilizing the LDA topic model, we were able to reveal the distributions of thematic sub-topics among the strategic documents. The topic modelling distributions were then used along with other document similarity measures as an input for the clustering of the strategic documents into groups. The results revealed three clusters of countries with a visible differentiation between the strategies of China and Japan on the one hand and the Scandinavian strategies (plus the German and the Luxemburgish) one on the other. The former promote technology and innovation-driven development plans in order to integrate AI with the economy, while the latter share a common view regarding the role of the public sector both as a promoter and investor but also as a user and beneficiary of AI, and give a higher priority to the ethical & safety issues that are connected to the development of AI.

CCS CONCEPTS

• **Applied computing** → **Computers in other domains** → Computing in government → *E-government*

• **Computing methodologies** → **Artificial intelligence**→ Natural language processing→ *Information extraction*

KEYWORDS

AI strategies, NLP, Automated Text Analysis, machine learning, topic modelling, document similarity

ACM Reference format:

Theodoros Papadopoulos, Yannis Charalabidis. 2002. What do governments plan in the field of artificial intelligence? Analysing national AI strategies using NLP techniques. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2020), 23-25 September 2020, Athens, Greece*, 12 pages. https://doi.org/10.1145/3428502.3428514

1. INTRODUCTION

Artificial intelligence is one of the most disruptive technologies of the twenty-first century: it is transforming industry and society, allowing for important changes at the global level by augmenting labor productivity and innovation, driving growth through intelligent automation, human-machine collaboration, and innovation diffusion. However, AI is not just another technology; it is a general-purpose technology that some analysts have likened to the combustion engine or electricity – which caused revolutions no country could afford to disregard [1]. Deep learning pioneer Andrew Ng compared AI to Thomas Edison's harnessing of electricity; *a breakthrough technology on its own, and one that once harnessed can be applied to revolutionize dozens of different industries* [2].

ICEGOV'20, September 23–25, 2020, Athens, Greece

ACM ISBN 978-1-4503-7674-7/20/09…$15.00
https://doi.org/10.1145/3428502.3428514

Moreover, as is the case with most disruptive innovations, the rapid progress in machine learning and the emerge of AI will inevitably trigger a new kind of geopolitical antagonisms. Multinational technology firms and industry leaders are scrambling to lead the development and use of artificial intelligence for the power and value it accrues, while Governments around the world are preparing to invest huge amounts of money in order to harness the potential economic and social benefits of its applications.

In order to prepare for the new geopolitical balance brought by artificial intelligence and to pave the way and facilitate the harness of AI, a growing number of countries have begun to develop national strategies on artificial intelligence. This publication uses NLP techniques on the raw text of twelve national AI strategies in order to facilitate their analysis and highlight key emerging topics. We analyzed all available national strategies at the time of writing this publication (Q3 2019), namely the strategies of the following countries: China, Denmark, Finland, France, Germany, India, Italy, Japan, Luxembourg, Mexico, Sweden, UK.

2. BACKGROUND

2.1. The geopolitics of AI

In March 2017, Canada became the first country to publish a national strategy to promote the use and development of AI [3]. Some months later, in October 2017, United Arab Emirates (UAE) became the first country in the world to institute a Ministry for Artificial Intelligence alongside the country's Strategy for AI [4]. Since then, several countries have devised national AI strategies or action plans to express their vision and prepare an enabling environment for the development and harness of AI technology. In the European Union, many member states have already published national strategies, while many others are working on preparing a strategic plan for the development and support of AI.

The growing number of countries with AI strategies, action plans (figure 1), and policies over the last three years demonstrates a clear recognition that AI must be a priority policy area. There is no other moment in the geopolitical history of the world, with so many governments releasing policy documents over a specific technology almost simultaneously.

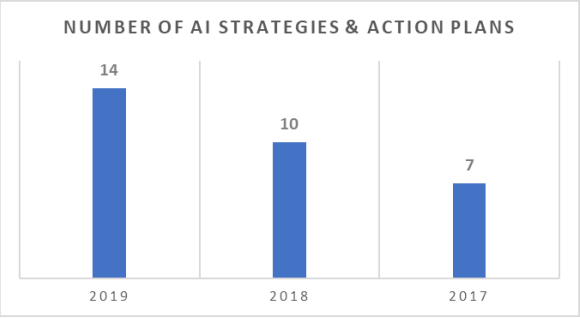

Figure 1: Number of national AI policies published per year

2.2. Text Analytics in Government

It is estimated that 80% of the world's data is unstructured [5], and a significant proportion of them are text-heavy. In order to find insights among unstructured text, text analytics techniques are employed to transform the text into data that can be used for further analysis. Text analytics is a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation [6]. One of the most important techniques used in text analytics is Natural Language Processing (NLP). Natural Language Processing employs methods both for Natural Language Understanding (NLU) and Natural Language Generation (NLG), which allows simulating the human ability το understand and create natural language text, e.g., to summarize information or infer topics from documents. Modern NLP techniques rely mostly on machine learning to derive meaning from human languages, using statistical inference to automatically learn rules through the analysis of large corpora of text.

The vast volume of unstructured text data filling into government agencies in both analog and digital form raises significant challenges for everyday operations, rulemaking, policy analysis, and citizen support. Text analytics techniques using NLP can provide the instruments needed to identify patterns and glean insights from all this data, allowing government agencies to improve operations, identify potential risks, work more efficiently, and improve public services.

2.3. Related Work

In this paper, we employ NLP and data mining technologies to discover knowledge by extracting, uncovering, and synthesizing information from a collection of national strategy documents on AI. Automated analysis of text has been employed in a wide variety of domains, but only a few target national strategies. This work is part of a recent but growing effort in applying advances in textual analysis to public policy research.

Our approach focuses on three applications of text analytics in public policy research. First, the extraction of keywords and summary for each document in order to gain insights about the main themes it covers. Second, topic modelling analysis in order to extract latent topics for each document, and third clustering of documents based on their semantic similarity.

Topic modelling techniques have been used to identify topics and sentiment in legislation and congressional debates [7,8]. Loni Hagein et al. [9] used topic modelling for the analysis of e-petition textual data to identify emergent topics of substantial concern to the public. Closest to our work is the work of Quinn et al., who used a topic model to examine the agenda in the U.S. Senate and inferred topical categories covered in a dataset of speeches from the Congressional Record [10]. Grimmer and King used clustering methods for an insightful conceptualization of political texts (press release data, State of the Union messages, and Reuters news stories data) [11]. Bousaills & Coan [12] and Farrell [13] use topic modelling to investigate climate change skepticism in reports and communications by think tanks and interest groups. Iglesias et al.

used NLP for the analysis of the communication policy (i.e., statements and minutes) of the Central Bank of Turkey [14]. They inferred that the communication policy of CBRT has been changing accordingly to the global economic conditions. Greene & Cross (2015) extracted latent thematic patterns in political speeches by developing a dynamic topic model to investigate how the plenary agenda of the European Parliament has changed over three parliamentary terms [15].

Several recent studies apply similarity analysis models to detect and cluster texts. Only a few of them, however, target semantic similarity detection for whole documents and very few target documents in the public policy research domain. Meade E. & Acosta, M. (2015), computed the "cosine similarity" between the "vector-space models" of consecutive statements of the Federal Open Market Committee in order to validate how persistent the content of the statements has been over time [16]. Ehrmann M. & Talmi J. (2019), computed the semantic similarity in central bank communication and market volatility [17]. LDA topics have rarely been used for semantic similarity. To the best of our knowledge, LDA topics distributions have not been used so far as a component to calculate the similarity between texts.

2.4. Methods for Document Understanding & Text Analytics

Following, we provide a brief introduction of the machine learning algorithms & NLP techniques used in our approach.

2.4.1. Word Embeddings

Since there are no specific features in text to analyze and run algorithms on it, we usually rely on word embeddings, a set of modelling and feature learning techniques in NLP where words or phrases from the plain text are mapped to vectors of real numbers. The technique of representing words as vectors has roots in the 1960s with the development of the vector space model (VSM).

The most straightforward approach, Bag of words (BoW) uses the whole pre-processed corpus to create a matrix in which the columns correspond to the tokens of the vocabulary, the rows to each document, and the value at each cell to the number of occurrence of a given token within a given document. The frequency of a word within a document is then calculated as the count of this word in the document divided by the total number of words in the document. This is also called the term frequency (TF) representation. BoW can be generalized such that each cell value does not necessarily show the exact term frequency, but a weight that represents a relevance measure of the term in the document. Some of the most popular weighting schemes are TF-IDF and BM25. The so-called Term Frequency-Inverse Document Frequency (TF-IDF), is computed by multiplying the term frequency (TF) with the inverse document frequency (IDF) which is a measure of the scarcity of a word across a collection of documents (calculated by dividing the total number of documents by the number of documents containing the term and then taking the logarithm of that proportion). Using TF-IDF, every word in a document is given a score, which increases proportionally with its frequency in the document and decreases if the word appears too often in the corpus. Traditional vector space models assign each word a singular dimension in their vector space, resulting in extremely sparse vector representations. Besides, merely treating words and phrases based on their frequency fails to consider word order and the semantics of the words. The term word embedding describes modern NLP techniques that encode the text into a lower-dimensional space as part of modelling its semantic meaning. Ideally, synonymous words and phrases end up with a similar representation in the new vector space. Word Embeddings have a long, rich history in NLP, but most versions depend in some way or another on the Distributional Hypothesis [18], which states that words that appear in the same contexts share semantic meaning.

One of the most popular approaches for distributional vectors is the word2vec model proposed by Mikolov et al. [19], which makes use of distributional or contextual information together with simple neural network models to obtain dense vector-space representations of words and phrases. Word2vec is a two-layer neural network, which uses distributional semantics to learn the correlation between words and their contexts. Two architectures are possible: continuous bag-of-words (CBOW), which is trained to predict words based on context words, and skip-gram, which takes a single word and tries to predict probabilities of other words being its context. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in the space. Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located close to one another [20]. The distance between word vectors also carries meaning, allowing to answer analogy questions using simple vector algebra, e.g., “King” - “man” + “woman” = “Queen” [19].

Word embeddings approaches have seen interesting developments recently, the most notable being FastText and ELMo (Embeddings from Language Models). FastText [21] was developed by the team of Tomas Mikolov and is an extension of the original word2vec framework. The main improvement is the inclusion of character n-grams, which allows computing word representations for words that did not appear in the training data (“out-of-vocabulary” words). Similar to fastText, ELMo [22] breaks the tradition of word embeddings by incorporating sub-word units, but ELMo also attempts to resolve the problem of words polysemy by computing contextual vector representations where the representation for each word depends on the entire context in which it is used. To achieve this, ELMo uses a deep bidirectional LSTM language model for learning words and their context. The deep Bi-LSTM architecture allows ELMo to learn more context-dependent aspects of word meanings in the higher layers along with syntax aspects in lower layers.

2.4.2. Text summarization & Keywords Extraction

Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning [23]. Text summarization can broadly be divided into two categories — Extractive and Abstractive Summarization. In extractive summarization, the generated

summary is a ranked selection of relevant sentences from a document or a collection of documents, while in abstractive summarization methods, advanced NLP techniques are used to generate an entirely new summary in which some parts may not even appear in the original text.

In our analysis, we use TextRank, a graph-based extractive summarization algorithm introduced by Rada Mihalcea and Paul Tarau [24]. TextRank is an unsupervised, domain, and language-independent algorithm that can be used both for keyword extraction and text summarization. The TextRank algorithm for keywords extraction uses a graph model to compute the rank of a word in an undirected, weighted word-graph using a slightly modified version of the PageRank algorithm [25]. To construct a word-graph for a given document, TextRank represents each term of the document as a node and then links two terms whenever both appear in a sliding window of a predefined size. Finally, it assigns the number of co-occurrences of the endpoints of an edge as a weight to the edge (term). When TextRank is used for document summarization, it applies the steps mentioned above on the sentence level linking from each sentence to all others or to the k-most similar sentences weighted by similarity.

2.4.3. Topic modelling

Topic modelling is an essential algorithm of text analytics. A topic model is a probabilistic model that discovers the main themes in a collection of documents. The basic idea is to treat the documents as mixtures of topics in the topic model, and each topic as a probability distribution of words. Topic modelling is an unsupervised machine learning technique, which means that it does not need training data in order to infer patterns and assign topics to a collection of documents. It resembles document clustering, but instead of 1-1 relationships between topic and cluster, in topic modelling a document can belong to many different clusters or topics.

In this paper, we are applying the Latent Dirichlet Allocation (LDA) topic modelling algorithm on the raw text of the twelve national strategies. LDA is a Bayesian-based model for text document collections represented by bags-of-words. The original paper [26] used a variational Bayes approximation of the posterior distribution, but alternative inference techniques such as Gibbs sampling and expectation propagation have been proposed [27,28]. The LDA model assumes that document collections have latent topics, in the form of a multinomial distribution of words, which is typically presented to users via its top-N highest probability words. Typically, only a small number of words are essential in each topic, and only a small number of topics are present in each document [29].

2.4.4. Documents Semantic Similarity

Semantic similarity is a widely used approach to the core problem of language understanding. The task of semantic similarity can be formulated at different levels of granularity ranging from word-to-word similarity to sentence-to-sentence similarity, document to-document similarity, or a combination of these such as word-to-sentence or sentence-to-document similarity [30]. Document to document similarity is useful for a wide variety of applications such as document clustering, recommender systems, plagiarism detection, and version control. In this paper, we use document similarity to cluster artificial intelligence strategies into a small number of groups. However, the definition of a pair of documents being similar or different is not always clear and varies typically with the actual problem setting. Accurate clustering of documents requires a precise definition of the closeness between a pair of documents, in terms of either their pairwise similarity or distance.

Whether using sparse or dense vectors, document similarities are computed by some function of the dot product between vectors. The cosine of two vectors—a normalized dot product—is the most popular such metric [31]. In the most straightforward approach, the cosine similarity is measured against vectors of the TF-IDF matrix and can be used to generate a measure of similarity between each document and the rest documents in the corpus. Other approaches use dense word embeddings such as the word2vec model, to compute a metric of similarity on the vectors representing each document. The most significant limitation of these approaches is their long computation time. To overcome this limitation, we introduce a different approach, based on the representations of texts as distributions over topics. Since the topic distributions for each document is a vector, we can compute the similarity between two documents by calculating the cosine similarity between their topic distributions.

2.4.5. Hierarchical Clustering

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters using an n × n vertex adjacency matrix containing distance values between pairs as input. There are two types of hierarchical clustering approaches i) Agglomerative: In which each node represents a single cluster at the beginning and the clusters are merged iteratively based on their similarities until all the elements belong to a single cluster, and ii) Divisive: In which initially, all nodes belong to the same cluster, and splits are performed recursively until each node forms its own cluster.

In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by the use of an appropriate metric (a measure of the distance between pairs of observations) such as the Euclidean or Manhattan distance, and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets. Some commonly used linkage criteria between two sets of observations are Maximum (complete-linkage clustering), Minimum (single-linkage clustering), or the increase in variance for the cluster being merged (Ward's criterion).

3. METHOD, IMPLEMENTATION AND RESULTS

3.1. Overview

Based on the raw text of twelve national AI strategies, this paper presents how NLP and data mining techniques can be applied to discover knowledge by extracting, uncovering, and synthesizing

information from a collection of documents. The methodology consists of a set of steps, each of which contributes to the overall analysis and knowledge discovery, allowing us to draw useful conclusions about the intentions and the priorities set by the strategic documents on AI. The main stages applied to the collection of text documents are shown in Fig.2

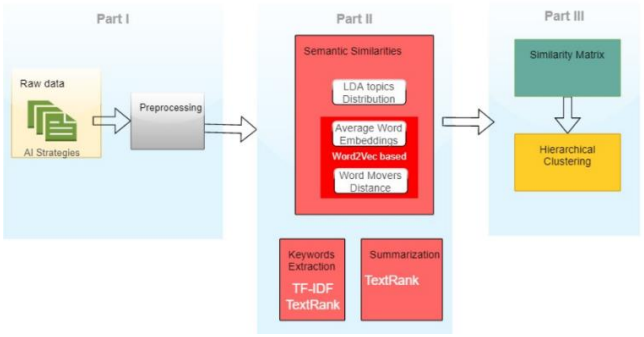

Figure 2: Overview of the methodology stages

Our approach pursues two primary goals. First, to cluster the AI strategic documents based on their semantic similarity and second, to reveal key themes inside each document. In order to cluster the national strategies based on their semantic similarity, we applied a novel approach by combining three different semantic similarity measures. The first utilizes LDA topic model distributions representing high-level information encoded in inferred topics, while the other two methods utilize dense vectors acquired via the google word2vec model. Besides, we applied the TF-IDF and TextRank algorithms to the corpus of documents in order to identify and rank relevant keywords and sentences and summarize the documents.

3.2. Coding Language, Environment and Source Code

The source code and the raw text data used for the analysis of the twelve national AI strategies are hosted online on github.com. The code has been implemented in Python 3.7 using Jupyter notebook as an execution environment and uses the following libraries: i) **gensim**, a Software Framework for Topic Modelling with Large Corpora, licensed under LGPL (https://github.com/RaRe-Technologies/gensim), ii) **spaCy**, a library for advanced Natural Language Processing in Python and Cython. Licensed under MIT license (https://github.com/explosion/spaCy), iii) **scikit-learn**, a Python module for machine learning, distributed under the 3-Clause BSD license (https://github.com/scikit-learn/scikit-learn), iv) **PDFMiner.six**, a tool for extracting information from PDF documents (https://github.com/pdfminer/pdfminer.six).

3.3. Documents Conversion and Pre-Processing

We retrieved the documents of twelve AI national strategies from the internet in pdf format. In order to extract the raw text from the pdf files, we prepared a python script that reads a folder with pdf files and converts them to plain text format (txt). For the conversion of each file, we used the open-source python package pdfminer.six.

Having the documents in a machine-readable format, we applied a series of pre-processing steps in order to prepare them for analysis. First, we converted the text into lower case letters because text analysis is case sensitive. We then used a set of regular expressions to remove content irrelevant to our analysis, such as special characters, links, punctuations, digits, and extra white spaces. To remove noise such as stop-words or terms which do not carry much weight in context to the text, we used the default dictionary of stop-words available from spaCy.

3.4. Corpus Analysis

Additional techniques such as phrase-detection (n-grams), tokenization, syntactic filtering, stemming, and lemmatization were employed to generate better-structured data suitable for corpus analysis. We used spaCy to segment the documents into word tokens (tokenization). Having pre-processed the documents in the form of a list of sequences of words (one sequence of words per document), we proceeded by transforming these word sequences into vectors of numerical features using the so-called bag-of-words representation, in which each document is represented by one vector where each vector element represents the times a word appears in the document. We then used gensim's phrases detection api to automatically detect and extract common bigrams and trigrams from the documents. Examples of such n-grams produced from the collection of documents are *artificial_intelligence*, *public_sector*, *machine_learning*. Before vectorization, we removed stop-words, filtered out parts of speech not useful for the analysis (keeping only verbs, adjectives, and nouns) and lemmatized the tokens using spaCy. To vectorize the documents, we first assigned a unique integer id to all terms appearing in the corpus using the gensim's Dictionary class, and then we converted the tokenized documents to a bag-of-n-grams corpus where the words in the documents were replaced with the respective id provided by this dictionary. Sorting the corpus by the total word count across all documents, we were able to extract the list with the most frequent words, as well as a list of most frequent bigrams across the corpus of documents (Table 1).

Table 1: top 15 term frequencies across the corpus

TF	TF (bi-grams)
(datum, 1623)	(artificial_intelligence, 1164)
(artificial_intelligence, 1164)	(public_sector, 158)
(technology, 1122)	(machine_learning, 106)
(research, 1072)	(public_administration, 91)
(development, 887)	(decision_making, 77)
(government, 680)	(public_authorities, 76)
(system, 626)	(private_sector, 66)
(service, 601)	(long_term, 59)

(develop, 571)	(large_scale, 50)
(work, 554)	(high_quality, 49)
(public, 547)	(start_ups, 48)
(sector, 525)	(computer_science, 47)
(digital, 517)	(business_models, 43)
(support, 514)	(deep_learning, 42)
(application, 499)	(higher_education, 39)

An overview of the extracted lists reveal the central theme of the documents (*artificial_intelligence*), their domain (*public_sector*, *public_authorities*, *public_administration*), as well as, various aspects such as the importance of data (datum), technology & research, the use of AI in decision making (*decision_making*) and the prevail of machine & deep learning approaches.

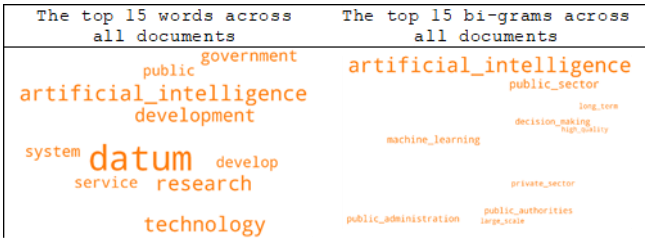

Figure 3: Word-cloud of the top 15 words across the corpus

3.5. Documents Summarization & Keywords Extraction

At the next stage of our analysis, we focused on revealing individual document's insights by extracting keywords and summarizing each strategic document.

To identify keywords for a specific document over the collection of documents, we used the measure of Term Frequency Inverse Document Frequency (TF-IDF). TF-IDF is a classic method for keyword extraction based on the so-called bag-of-words model, but unlike the regular bag-of-words, TF-IDF downweighs tokens (words) that frequently appears across documents, intending to reflect the importance of a word to a document in a collection or corpus. For our analysis, we computed the TF-IDF matrix using gensim. The matrix is in the form of an array of vectors, where each vector represents a document. The vector contains a list of weights for each unique word in the dictionary with the TF-IDF value if the word is in the document, or 0.0 otherwise. Table 2 presents the top 10 words of each document along with their weight value, as extracted from the TF-IDF matrix.

Table 2: top 10 terms per document (TF-IDF)

Country	Top 10 terms per document (TF-IDF)
China	('theory', 0.464), ('intelligent', 0.245), ('construct', 0.216), ('comprehensively', 0.212), ('swarm_intelligence', 0.199), ('sensing', 0.185), ('platform', 0.147), ('quantum', 0.133), ('equipment', 0.13), ('human_machine', 0.13)
Denmark	('ethical_principles', 0.264), ('public_authorities', 0.258), ('business_community', 0.207), ('competence', 0.166), ('freely_available', 0.151), ('growth_fund', 0.132), ('outset', 0.132), ('priority_areas', 0.131), ('healthcare', 0.121), ('utility', 0.12)
Finland	('utilisation', 0.542), ('utilise', 0.244), ('accelerator', 0.193), ('labour_market', 0.181), ('actor', 0.148), ('economic_affairs', 0.14), ('filter', 0.121), ('competence', 0.114), ('steering_group', 0.109), ('excellent', 0.105)
France	('occupation', 0.202), ('exception', 0.162), ('public_procurement', 0.133), ('girl', 0.127), ('public_authorities', 0.121), ('fact', 0.107), ('female', 0.103), ('european', 0.1), ('frequently', 0.098), ('chair', 0.092)
Germany	('mittelstand', 0.192), ('standardisation', 0.172), ('government_wants', 0.148), ('dialogue', 0.142), ('european', 0.138), ('competition_law', 0.137), ('observatory', 0.126), ('consultation_process', 0.126), ('social_partners', 0.118), ('civil_society', 0.118)
India	('marketplace', 0.349), ('annotation', 0.254), ('adoption', 0.193), ('poor', 0.184), ('farmer', 0.146), ('smart_cities', 0.133), ('value_chain', 0.127), ('intervention', 0.1), ('incentivise', 0.099), ('insight', 0.095)
Italy	('public_administration', 0.371), ('inequality', 0.186), ('citizen', 0.178), ('task_force', 0.12), ('guarantee', 0.109), ('accompany', 0.109), ('imagination', 0.099), ('precisely', 0.099), ('interpret', 0.099), ('human_beings', 0.097)
Japan	('utilization', 0.444), ('realize', 0.253), ('utilize', 0.234), ('industry_academia', 0.186), ('human_resources', 0.186), ('foster', 0.162), ('maintenance', 0.16), ('attachment', 0.159), ('personnel', 0.156), ('relevant_ministries', 0.147)
Luxembourg	('human_centric', 0.335), ('digitalization', 0.26), ('vision', 0.235), ('powered', 0.144), ('personalized', 0.137), ('analyze', 0.115), ('cross_border', 0.115), ('augment', 0.112), ('topic', 0.109), ('laboratory', 0.101)
Mexico	('automatable', 0.166), ('oxford_insights', 0.166), ('interview', 0.165), ('recommend', 0.151), ('proportion', 0.147), ('score', 0.147), ('civil_society', 0.146), ('acronym', 0.142), ('contracting', 0.142), ('table', 0.139)
Sweden	('safe_secure', 0.387), ('county', 0.25), ('higher_education_institutions', 0.205), ('harness', 0.17), ('ethical_principles', 0.167), ('testbed', 0.167), ('realise', 0.132), ('defence', 0.129), ('welfare', 0.129), ('manipulate', 0.129)
Uk	('fibre', 0.177), ('cluster', 0.171), ('catapult', 0.147), ('fellowship', 0.147), ('visa', 0.147), ('grand_challenges', 0.144), ('case_study', 0.137), ('thrive', 0.126), ('prestigious', 0.118), ('dame', 0.118)

In order to gain insights about the documents, we used TextRank to summarize the documents. Table 3 presents the summary of the national strategies of Denmark and Luxemburg as produced by the TextRank algorithm. The summarized version allows us to overview the major points of the original document in a fast but representative way.

Table 3: summaries for Denmark & Luxemburg

Summary **(Denmark)**
Researchers in Denmark will research artificial intelligence through basic research and more application-oriented research to pave the way for the development of useful technological solutions for the individual, businesses and the public sector. The goals of the government are that: Denmark is among the best in the world at exploiting data and new business models based on responsible development and use of artificial intelligence. With **a strong focus on ethics**, better use of data, and the importance of competences and research, the government is also sowing the seeds for Denmark to play a role in developing artificial intelligence in the long term. Technical and organisational solutions should be developed that **support ethically responsible development and use of artificial intelligence** in order to achieve the greatest possible progress for society, e.g. by contributing to better public-sector services and to growth in the business community.
Summary **(Luxembourg)**

This policy vision is built on Luxembourg's ambitions as a digital front-runner:
Ambition #1: - To be among the most advanced digital societies in the world, especially in the EU
Ambition #2: - To become a data-driven and sustainable economy
Ambition #3: - To **support human-centric AI development** It is possible that the country's opportunities to develop ground-breaking fundamental AI research are limited.
A **human-centric** approach requires a level playing field among the diverse stakeholders investing, working and living in Luxembourg – ensuring they all benefit fully from new technologies and the nation's digitalization.
Luxembourg's public research actors are actively involved in developing state-of-the-art technologies, paving the way for smart cities and the technological applications that go with it (smart connectivity, smart mobility, smart health, smart living, etc.) to improve urban life through integrated and sustainable solutions.

3.6. Extraction of Topics

The objective of topic models is to extract the underlying topics from a given collection of text documents. The LDA Topic modelling technique was applied to the pre-processed bag of words for each strategy document to generate the cluster of topics and the terms belonging to each cluster topic. The two main inputs to the LDA topic model are the dictionary and the corpus. For the dictionary, we decided to exclude word with low Document frequency (DF), setting the low-DF threshold at 1, thus only excluding singletons- spelling-errors and very rare terms that would not contribute to topic modelling. The corpus is the Term Document Frequency and is a mapping of word to word frequency for each strategic document.

In addition to the corpus and dictionary, LDA requires to provide the number of topics as well. In order to determine the best value for the number of topics, we run several experiments in which we calculated the coherence value of each model. Topic coherence is one of the primary metrics used to estimate the number of topics. We used both UMass and c_v measure to calculate the coherence score of various LDA models, and we decided to proceed with four topics.

Table 4 presents the topic distributions and the dominant topic (values in bold) for each document. We have excluded contributions that are less than 1% from the results (substituted with zero). Figure 4 presents the count and weight of the top 8 tokens for each topic. Note that topics are generally not mutually exclusive but often share common words. However, the weighting of each word is different between topics (Figure 4).

Table 4: dominant topic and topics distribution per strategy

Country	Dominant Topic	Topic0	Topic1	Topic2	Topic3
China	2	0	0.435	**0.554**	0
denmark	0	**0.968**	0	0.029	0
finland	0	**0.493**	0	0.085	0.422
france	1	0.39	**0.445**	0	0.165
germany	0	**0.505**	0.183	0.212	0.1
india	1	0.395	**0.396**	0.087	0.122
italy	0	**0.435**	0.369	0.056	0.14
japan	2	0.293	0.235	**0.39**	0.083
luxembourg	0	**0.479**	0.286	0.109	0.126
mexico	0	**0.448**	0.35	0.065	0.136
sweden	0	**0.627**	0.107	0.177	0.089
uk	0	**0.537**	0.269	0.096	0.099

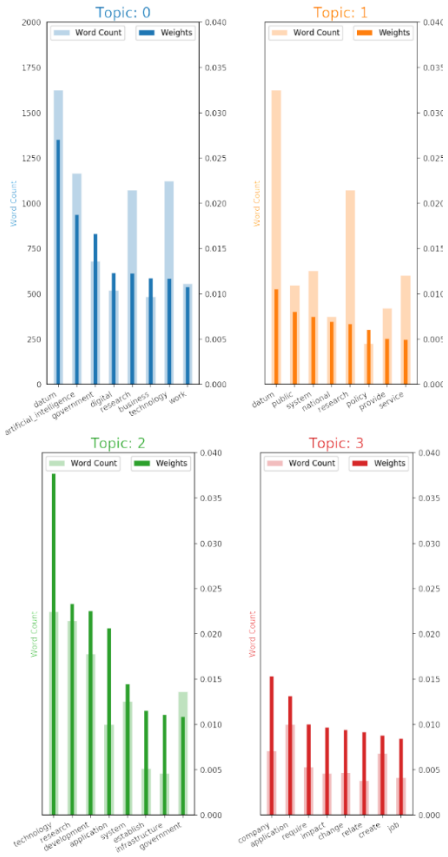

Figure 4: Word-count & weights for topics (top-8 tokens)

3.7. Documents Similarities

In order to calculate the similarity between documents, we combined three different approaches. The first one is based on topic modelling results from the previous stage and is implemented by calculating the cosine similarity over the topic distributions for each pair of documents. The rest two approaches use a pre-trained Word2vec model in order to calculate the cosine similarity over the word vectors for each pair of documents using two different measures: The Average Word Embeddings (AWE) and Word Mover's Distance (WMD). Table 5 presents the results for each method. The content of each cell shows the three calculated similarities in the upper part of the cell (LDA topic distributions, Word2vec AWE, Word2vec WMD) as well as the

mean of these values in the lower part of the cell. Highlighted cells indicate pairs of documents with maximum mean similarity.

3.7.1. First Method: cosine similarity over topic distributions

Cosine similarity is a measure of similarity that measures the cosine of the angle between two vectors projected in a multi-dimensional space. For our analysis, we calculated the cosine similarity between all the pairs of vectors containing the topics distribution of each strategic document. The intuition behind this is that two documents with similar topics distributions will be more similar than two documents with very different topic distributions.

The results on table 5, verify that pairs of countries having balanced topics distribution will also have a high cosine similarity value (first value in the upper part of each cell), while pairs of countries with uneven distributions will have low cosine similarity.

3.7.2. Second Method: Average word embeddings

The second method exploits the information contained in word vectors. We used a pre-trained Word2vec model from google (word2vec-google-news-300) that includes word vectors for a vocabulary of 3 million words and phrases trained on roughly 100 billion words from a Google News dataset. The vector length is 300 features. Because the google model is not trained on the same corpus, we first removed out-of-vocabulary words from our corpus. For each document, we computed the mean of its word's vectors so that each document is represented by a 300-dimensional vector, and then we used the pairs of those vectors to compute the cosine similarity between documents. The second value in the upper part of each cell in Table 5 contains the calculated similarity using this method.

3.7.3. Third Method: Word Mover's Distance (WMD)

The Word Mover's Distance method was introduced in the article "From Word Embeddings to Document Distances" by Matt Kusner et al. [32]. It is inspired by the "Earth Mover's Distance" and employs a solver of the "transportation problem". WMD seeks to calculate the minimum cumulative "traveling distance" that words from a reference document need to travel to match words from another document. Because the result is a distance measure (values closer to zero indicate similar documents), we subtracted them from 1 in order to transform them into a similarity measure. The third value in the upper part of each cell in Table 5 contains the calculated similarity using this method.

Table 5: Document Similarities based on LDA topic distributions, W2V Centroid & W2V WMD

	CN	DK	FI	FR	DE	IN	IT	JP	LU	MX	SE	UK
CN	-	0, .92, .92 .615	0, .91, .92 .611	.25, .90, .94 .698	.39, .93, .94 .752	.39, .95, .95 .760	.32, .92, .91 .715	.86, .96, .94 **.919**	.39, .95, .94 .759	.28, .94, .94 .721	.14, .94, .93 .669	.21, .94, .93 .695
DK	0, .92, .92 .615	-	.48, .99, .97 .815	.25, .98, .95 .728	.75, .99, .95 .896	.33, .98, .94 .749	.51, .98, .93 .806	.27, .97, .92 .720	.61, .98, .94 .843	.42, .97, .94 .775	.90, .97, .95 **.940**	.48, .97, .94 .795
FI	0, .91, .92 .611	.48, .99, .97 .815	-	.40, .99, .95 .778	.70, .98, .95 .877	.31, .98, .94 .739	.58, .98, .92 .828	.41, .97, .91 .764	.56, .97, .94 .822	.39, .97, .93 .763	.73, .96, .95 **.879**	.30, .96, .93 .729
FR	.25, .90, .94 .698	.25, .98, .95 .728	.40, .99, .95 .778	-	.74, .99, .97 .902	.98, .98, .96 **.974**	.95, .99, .94 .961	.65, .97, .94 .851	.90, .98, .96 .945	.98, .98, .95 .969	.59, .97, .96 .837	.95, .96, .95 .951
DE	.39, .93, .94 .752	.75, .99, .95 .896	.70, .98, .95 .877	.74, .99, .97 .902	-	.78, .99, .96 .910	.92, .99, .93 .946	.77, .98, .94 .895	.96, .99, .96 **.967**	.83, .98, .95 .918	.95, .98, .97 .964	.80, .97, .96 .909
IN	.39, .95, .95 .760	.33, .98, .94 .749	.31, .98, .94 .739	.98, .98, .96 .974	.78, .99, .96 .910	-	.95, .98, .95 .960	.73, .99, .96 .891	.93, .99, .97 .962	.99, .99, .97 **.981**	.62, .98, .96 .850	.97, .98, .97 .971
IT	.32, .92, .91 .715	.51, .98, .93 .806	.58, .98, .92 .828	.95, .99, .94 .961	.92, .99, .93 .946	.95, .98, .95 .960	-	.74, .97, .93 .879	.99, .98, .94 **.970**	.97, .98, .95 .968	.80, .97, .94 .905	.94, .96, .95 .947
JP	.86, .96, .94 **.919**	.27, .97, .92 .720	.41, .97, .91 .764	.65, .97, .94 .851	.77, .98, .94 .895	.73, .99, .96 .891	.74, .97, .93 .879	-	.78, .98, .95 .902	.68, .97, .95 .866	.54, .97, .94 .816	.59, .96, .94 .833
LU	.39, .95, .94 .759	.61, .98, .94 .843	.56, .97, .94 .822	.90, .98, .96 .945	.96, .99, .96 .967	.93, .99, .97 .962	.99, .98, .94 **.970**	.78, .98, .95 .902	-	.95, .99, .96 .966	.85, .98, .95 .928	.93, .98, .96 .955
MX	.28, .94, .94 .721	.42, .97, .94 .775	.39, .97, .93 .763	.98, .98, .95 .969	.83, .98, .95 .918	.99, .99, .97 **.981**	.97, .98, .95 .968	.68, .97, .95 .866	.95, .99, .96 .966	-	.70, .98, .95 .874	.99, .98, .96 .975
SE	.14, .94, .93 .669	.90, .97, .95 .940	.73, .96, .95 .879	.59, .97, .96 .837	.95, .98, .97 **.964**	.62, .98, .96 .850	.80, .97, .94 .905	.54, .97, .94 .816	.85, .98, .95 .928	.70, .98, .95 .874	-	.70, .96, .95 .869
UK	.21, .94, .93 .695	.48, .97, .94 .795	.30, .96, .93 .729	.95, .96, .95 .951	.80, .97, .96 .909	.97, .98, .97 .971	.94, .96, .95 .947	.59, .96, .94 .833	.93, .98, .96 .955	.99, .98, .96 **.975**	.70, .96, .95 .869	-

3.8. Clustering based on pair-wise similarities

To cluster the twelve documents, we first converted the mean similarity matrix in a distance matrix (by substituting one from all values), and then we applied agglomerative hierarchical clustering analysis using the Euclidean distance as a metric and Ward's criterion as a measure of the distance between sets of observations. The clustering results in the form of a dendrogram (Figure 5) provide an easy-to-interpret view of the clustering structure. The vertical axis (heights of each cluster) reflects the distance or dissimilarity between the clusters.

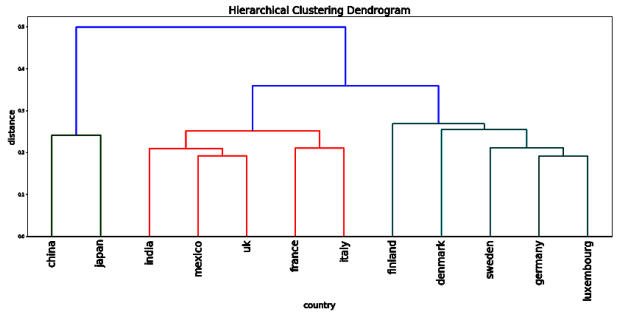

Figure 5: Hierarchical clustering (dendrogram)

Looking at the dendrogram, one can see the three clusters as three branches consisting of China and Japan (in Cluster 1), Denmark, Finland, Germany, Luxemburg, and Sweden (in Cluster 2) and the rest of the countries in Cluster 3. Clusters 2 and 3 are merging in a lower height (at about 0.35) while cluster 1 merges with clusters 2 and 3 in a much higher height (at about 0.5), capturing the differentiation between the strategies of China & Japan and the rest of the countries. If we choose a 2-clusters representation, we would have China and Japan in one cluster with the rest of the countries in the other one.

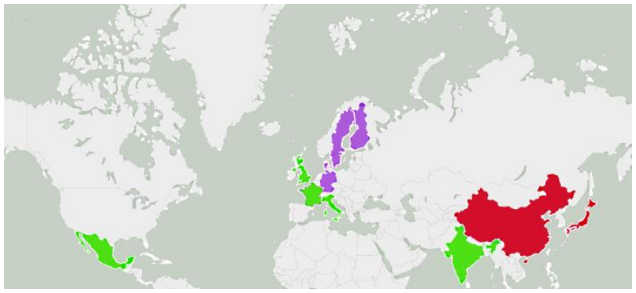

Figure 6: clustering results on map (3 clusters)

4. DISCUSSION OF RESULTS

4.1. Countries Similarities

China and Japan (cluster 1) have a similar approach to AI, focusing more on national initiatives and large-scale projects. A review of the results of the analysis reveals a technology-based approach, with references to cutting edge technologies, big data, smart cities, and machine learning applications. The vision in this cluster seems to be an innovation-driven development strategy to accelerate the deep integration of AI with the economy. An excerpt from China's strategy, referring to the Basic Principles of the strategy, states: "***Technology-Led***. *Grasp the global development trend of AI, highlight the deployment of forward-looking research and development, explore the layout in key frontier domains, long-term support, and strive to achieve transformational and disruptive breakthroughs*". China's strategy also clearly refers to the possible use of AI for automatic decision making as well as, to the integration of AI into government services.

Countries in cluster 2 (Denmark, Finland, Germany, Luxemburg, and Sweden) share a common view regarding the role of the public sector both as a promoter and investor but also as a user and beneficiary of AI. Most of those strategies envision the involvement of the public sector in setting up shared sector platforms, which will provide secure and tailored access for the various participants of different ecosystems (researchers, companies, public authorities). The strategies in this group support that the active promotion of AI applications in public sector activities can play a significant role in how the public sector succeeds in responding to significant social challenges of the future. There is also a greater coverage on ethical and safety issues, reflecting the higher priority given within the EU to these issues.

UK and Mexico in cluster 3, have very high pair-wise similarities between them (as can also be seen from the minimum height between their nodes on the HAC dendrogram). A possible explanation for the high similarity of the UK-Mexico pair could be that Mexico's strategy was commissioned by the British Embassy in Mexico, funded by the UK's Prosperity Fund, and developed by Oxford Insights, a UK-based consultancy firm.

4.2. Key Findings

4.2.1. Importance of Data

Data (Datum) is the most frequent word in the corpus (Table 1) and a key token in most of the extracted topics (figure 4). The most common reference to data relates to their recognition as a key ingredient for the progress and development of AI. Querying the corpus, we were able to extract a set of sentences with high weighting that recognize data availability, quality, openness, and governance as prerequisites for the development of the AI. Table 7 presents some related findings for the terms "data quality" & "open data". Most strategies recognize data as the cornerstone of AI development and promote actions for better access to public sector data. Data governance is a key term in the documents, highlighting the need for high data quality throughout their complete lifecycle.

Table 7: sentences for data quality, access & openness

data / data quality	**open data / access to data**
data quality, data security and data curation are horizontal issues and, therefore, targeted funding needs to be provided across all sectors	this opening-up could take one of two forms access to this **data** by public authorities alone, in order to feed into **a** public **data** platform, for example or wider access **open data** which would be **open** to other economic stakeholders.
data annotation marketplace	we will explore providing targeted funding for **open** training **data** sets that are compatible with **data** protection rules.
data infrastructure ai is powered by data.	the provision of **open** government **data** for unrestricted further use is to be expanded.

data the cornerstone of **AI**	it is also important to **open** privately-owned **data** where necessary.
data and platform governance is regularly underestimated, both as regards collection what needs to be collected and how and data management over time	**a** second area of discussion is the management and research of **data** published on the web in the form of linked **open** data.
challenges include **data** usage without consent, risk of identification of individuals through data, **data** selection bias and the resulting discrimination of ai models, and asymmetry in **data** aggregation.	the national digital strategy promotes the publication of **open data**
data availability and **quality** is vital to ai research and to the usefulness of ai programmes.	better access to public-sector **data**.
data that is enriched with comments and metadata.	each administration is required to issue **open data** to contribute to the enhancement of public information assets, in line with international and national **open data** policies.
data production supervision of public dataset quality management and promotion of **data** access regulation of **data** usage.	improved access to **data** outside Denmark for danish businesses and researchers

4.2.2. *Importance of ethical & AI safety issues*

A second important finding revealed by examining the results is that all the AI strategies focus, to a greater or lesser extent, on the ethical issues related to the development of artificial intelligence and emphasize the importance of safety and safeguards in AI-based decision making. This emphasis is more apparent in the documents of cluster 2 (Denmark, Finland, Germany, Luxemburg, Sweden).

The term *ethical_principles* is the most relevant for Denmark's strategy according to the TF-IDF algorithm, while for Sweden, *ethical_principles* & *safe_secure* are amongst the top-10 relevant terms. Most strategies acknowledge that in the future, the use of AI technology is expected to raise new legal and ethical issues, and several of them are proposing the establishment of an ethics supervision agency, council or commission.

As history suggests, when humanity is confronted with new disruptive or controversial technologies such as nuclear power and genetically modified foods, safety issues are arising and introduced early on to the debate. Given that AI algorithms in decision-making can have significant societal impact and will affect horizontally our lives, the emergence of ethical and safety issues in the policy documents is reflecting the broader reservations about the capitulation of decision-making, an exclusively human responsibility till now, to artificial intelligence algorithms.

Table 8: ethics & safety terms frequency

ethics-related terms	safety-related terms
('ethic', 124)	('safety', 42)
('ethical', 104)	('safe', 26)
('ethical_issues', 17)	('safeguard', 20)
('ethical_principles', 16)	('safe_secure', 8)
('ethically', 14)	
('ethics_commission', 9)	
('ethical_questions', 8)	
('ethical_considerations', 8)	

4.2.3. *Skills & Education*

Education and skills are also a key pillar of most of the AI strategies examined. One the one hand, the harness of AI demands a highly specialized workforce with new skills and very high levels of competence in several areas, and on the other hand, AI automation will replace existing jobs increasing the demands for up-skilling and retrain of the existing workforce. Table 9 presents the skills -related terms frequency in the corpus of the AI strategies.

Table 9: skills & education terms frequency

skills-related terms	education-related terms	Training-related terms
('skill', 290)	('education', 249)	('training', 216)
('skilled', 41)	('higher_education', 39)	('train', 83)
('highly_skilled', 6)	('educational', 37)	('vocational_training', 26)
('upskill', 4)	('educate', 18)	('trainer', 6)

With the increased demand for AI talent, there is now a significant shortage of people with skills and knowledge to carry out major AI research projects across different industries. Under such circumstances, salaries are skyrocketing, leading to an AI brain drain and drawing scientists and researchers away from the academic institutions where artificial intelligence research is taking place. Besides, a similar flow occurs from low-income to high-income countries. The need for attracting skill potential in Artificial Intelligence and the importance of avoiding brain drain is particularly evident in European Strategies. For instance, the German AI Strategy states, "We *will make the task of harnessing the domestic and European skills potential a priority. The third leg of our skills strategy is about tapping the international skills potential* ", as well as, "*Vocational training and education need to be adapted to the changing requirements linked to the digital transformation and, in this context, AI".*

4.2.4. *The race for leading in AI*

Another important theme that has also been revealed through the analysis of the twelve AI strategies is the ambition of many countries to lead in the AI race. By querying the processed corpus of the AI strategies for tokens stemming from "lead", we had the following results: *leader* (51), *leadership* (43), *world_leading* (15), *world_leader* (11), *leading_role* (9), *global_leader* (7). By extracting relative sentences from the corpus, we were able to identify many references to AI world-leading ambitions.

For instance, The German Strategy on AI states "*We want Germany to build on its strong position in Industrie 4.0 and to*

*become **a world leader** in AI applications*", as well as "*We want Germany to build upon its very good position in AI research, … and **to become a world leader** in this area*". The Swedish Strategy also states that "*the Government's goal is to **make Sweden a leader** in harnessing the opportunities that the use of AI can offer, with the aim of strengthening Sweden's welfare and competitiveness*". Finland's strategy describes the vision to "*make Finland **a leading country** in the application of artificial intelligence*", while the UK Strategy states that "*A revolution in AI technology is already emerging. If we act now, **we can lead it** from the front. …. Together, we **can make the UK a global leader** in this technology that will change all our lives*". In Mexico's strategy, we read that "*Mexico should aim at **being a global leader** in AI and digitalisation as a way of promoting development, both social and economic*". China has set an explicit goal at the highest level of government to make itself the global leader in AI "*by 2030, China's AI theories, technologies, and applications should **achieve worldleading levels**, making China the world's primary AI innovation center*".

The ambitions for the lead in AI development can also be confirmed by political statements and speeches outside of the strategies. Emmanuel Macron declared: "*I want **France to become a leader in AI**, we have the capacities, we must create the proper conditions*" (Emmanuel Macron speech at the prestigious Collège de France). However, it is the Russian president, Vladimir Putin, who has most clearly expressed the dominant role of AI in the new era's geopolitics. Putin, in his official speech for the inauguration of the Russian school year in September 2017, emphasized that "*whoever became **the leader in the field** would rule the world*".

The aforementioned statements reinforce the view that governments all over the world have already started to see AI as the key differentiating technology of our era, and a new geopolitical race is about to start. A narrative has begun to emerge in both policy and public context [34], arguing that the AI race will lead to geopolitical tensions and rivalries both between and across countries and multinational companies. Those tensions will shape the 21st century and will probably reinforce nationalisms around the world. As the authors correctly argue in [33] "*As was the case with the space race during the Cold War-era, the increasing development and imminent deployment of AI by states and, in civilian and military domains will threaten the current global order – even reshaping how states and businesses operate and succeed*".

The race for technological superiority reflects a perception that leadership in AI could secure clear infrastructural and economic advantages to frontrunners. The likelihood of such a race, however, poses many adverse effects. A struggle for AI supremacy would not be conducive for international cooperation and would undermine the efforts to ensure the proper safety precautions that such technology will require. Such considerations become particularly crucial if AI approaches superintelligence or progress towards the development of lethal autonomous weapons. Besides, a prolonged AI race narrative poses a specific capacity for fabricating and sustaining antagonistic relations and nationalisms between groups and countries. In the case of the race for technological advantage, it encourages people to see people from competing countries as threats or even enemies [34], intensifying the already existing nationalisms. Last but not least, if the AI race finally were won, the concentration of power in the hands of whatever group, firm, or country possesses AI superiority will allow the already powerful to consolidate their power further obliterating many practical advantages of democracy and eroding the ideals of non-discrimination and equality. If this struggle for supremacy in AI escalates, and if authoritarian regimes or the growing segment of the super-rich elite were to emerge victoriously, then a dystopian future with reduced human autonomy, inequality, and authoritarian practices would not seem unlikely.

In this context, it is noteworthy that individual European member states adopted the rhetoric for supremacy in artificial intelligence. Our analysis revealed that most of the European strategies envision a world-leading role with very few references to the pan-European efforts for cooperation. While the European Commission is trying to promote pan-European initiatives and collaborations, member states remain attached to the era of national autonomy, pursuing national superiority in fields that are considered critical for obtaining or maintaining comparative advantages. For European member states to seriously pursue influence on the global development of AI, instead of undermining the Commission's initiatives to coordinate AI efforts, should recognize that only by joining forces, the critical mass for serious AI investments can be reached and pursue cooperation at a European level. Only at this level can the European states enforce regulatory and safety standards for an "ethical AI" that the rest of the world will have to follow.

5. CONCLUSIONS AND FUTURE WORK

Although AI has been discussed for decades, only recently has it received serious and sustained attention from governments. The recent advantages in the AI filed especially in machine learning along with other technological advantages that are giving machines access to a vast amount of data or the ability to act in the real world, have demonstrated the potential of this technology, creating a sense of immediacy about seizing the AI opportunity to governments and multinational organizations. In a brief period of three years, between 2017 and 2019, over 30 countries have developed national AI strategies or action plans. Our approach highlights insights by applying and combining NLP techniques in a corpus of national AI strategies. The corpus-wide analysis revealed key findings such as the importance of the data ecosystem for the development of AI, the increasing considerations about ethical and safety issues, as well as, the imminent global race for AI dominance.

In order to cluster the national strategies based on their semantic similarity, we applied a novel approach by combining three different similarity measures. The first utilizes LDA topic model distributions representing high-level information encoded in inferred topics, while the other two methods utilize dense vectors acquired via the google word2vec model. The results revealed that there is a visible differentiation between China's and Japan's strategies, which are promoting technology and

innovation-driven development plans in order to integrate AI with the economy and the rest of the strategies examined. Scandinavian and German strategies tend to acknowledge the importance of public sector in the development of AI, as well as that in order to remain competitive in the global race for AI, the European member states need to focus more on the development of AI-relevant skills, and shared resources including data, and a safe regulatory environment.

Although we built our experiments using texts in English, our approach is language-agnostic because it depends only on vector representations of the text and can thus be applied to archives in a multitude of languages. What our methodology currently lacks is the theoretical framework that can support the selection of the weights for each similarity measure. The selection of the current formula (mean of the three similarity measures) is mainly intuitive and based on an empirical approach, without providing solid theoretical foundations of its efficacy.

As future work, we would like to test two improvements in our approach, (1) the automation of the stop-words generation based on the relevance of each word to the corpus and (2) the combination of more than three similarities measures to calculate the overall similarity, possibly using various word embeddings models.

REFERENCES

[1] European Council on Foreign Relations. 2019. Machine politics: Europe and the AI revolution. Retrieved September 14, 2019 from https://www.ecfr.eu/publications/summary/machine_politics_europe_and_the_ai_revolution#

[2] Lee, K. F. 2018. AI superpowers: China. Silicon Valley, and the New World Order. Houghton Mifflin Harcourt.

[3] Canadian Institute for Advanced Research. 2017. Pan-Canadian Artificial Intelligence Strategy Overview, Report, Canadian Institute for Advanced Research (CIFAR), Toronto, Canada

[4] ArabianBusiness.com. 2019. UAE appoints first Minister for Artificial Intelligence. Retrieved 8 December 2019, from https://www.arabianbusiness.com/politics-economics/381648-uae-appoints-first-minister-for-artificial-intelligence

[5] John Gantz and David Reinsel. 2011. Extracting Value from Chaos. Retrieved from https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[6] Wikipedia.2019. Text mining. Retrieved from https://en.wikipedia.org/wiki/Text_mining

[7] Thomas, Matt & Pang, Bo & Lee, Lillian. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts.

[8] Hillard, Dustin & Purpura, Stephen. 2007. An Active Learning Framework for Classifying Political Text.

[9] Loni Hagen, Ozlem Uzuner, Christopher Kotfila, Teresa M. Harrison, and Dan LaManna. 2015. Understanding citizens' direct policy suggestions to the federal government: A natural language processing and topic modeling approach. In Proceedings of the Annual Hawaii International Conference on System Sciences. DOI:https://doi.org/10.1109/HICSS.2015.257

[10] Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. Am. J. Pol. Sci. (2010). DOI:https://doi.org/10.1111/j.1540-5907.2009.00427.x

[11] Grimmer, Justin & King, Gary. 2011. General Purpose Computer-Assisted Clustering and Conceptualization. Proceedings of the National Academy of Sciences of the United States of America. 108. 2643-50. 10.1073/pnas.1018067108.

[12] Boussalis, Constantine & Coan, Travis. 2016. Text Mining the Signals of Climate Change Doubt. Global Environmental Change. 36. 89-100. 10.1016/j.gloenvcha.2015.12.001.

[13] Justin Farrell. 2016. Corporate funding and ideological polarization about climate change. Proc. Natl. Acad. Sci. U. S. A. (2016). DOI:https://doi.org/10.1073/pnas.1509433112

[14] Joaquin Iglesias & Alvaro Ortiz & Tomasa Rodrigo, 2017. How do the EM Central Bank talk? A Big Data approach to the Central Bank of Turkey".Working Papers 17/24, BBVA Bank, Economic Research Department

[15] Derek Greene and James P. Cross. 2017. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. Polit. Anal. (2017). DOI:https://doi.org/10.1017/pan.2016.7

[16] Meade, E. E., & Acosta, M. 2015. Hanging on every word: Semantic analysis of the FOMC's postmeeting statement (No. 2015-09-30). Board of Governors of the Federal Reserve System (US).

[17] Ehrmann, M., & Talmi, J. 2019. Starting from a blank page? Semantic similarity in central bank communication and market volatility. Journal of Monetary Economics.

[18] Zellig S. Harris .1954. Distributional Structure, WORD, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. (January 2013). Retrieved October 19, 2019 from http://arxiv.org/abs/1301.3781

[20] Wikipedia. 2010. Word2vec. Retrieved from https://en.wikipedia.org/wiki/Word2vec

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146

[22] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. NAACL

[23] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. Retrieved September 16, 2019 from http://arxiv.org/abs/1707.02268

[24] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. Proc. EMNLP (2004). DOI:https://doi.org/1.3115/1219044.1219064

[25] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. World Wide Web Internet Web Inf. Syst. (1998). DOI:https://doi.org/1.1.1.31.1768

[26] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 2003. DOI:https://doi.org/1.1016/b978-0-12-411519-4.00006-9

[27] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. Proc. Natl. Acad. Sci. U. S. A. (2004). DOI:https://doi.org/1.1073/pnas.0307752101

[28] Thomas Minka and John Lafferty. 2002. Expectation-Propagation for the Generative Aspect Model. Uncertain. Artif. Intell. DOI:https://doi.org/ISBN 1-55860-897-4

[29] Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. 2010. Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 100-108).

[30] Niraula, N., Banjade, R., Ştefănescu, D., & Rus, V.2013. Experiments with semantic similarity measures based on lda and lsa. In International Conference on Statistical Language and Speech Processing (pp. 188-199). Springer, Berlin.

[31] D Jurafsky and J H Martin. 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Prentice Hall

[32] Matt J Kusner, Yu Sun, Nicholas I Kolkin, and Kilian Q Weinberger.2009. From Word Embeddings to Document Distances.

[33] Samir Saran, Nikhila Natarajan, and Srikumar Madhulika. 2018. In Pursuit of Autonomy: AI and National Strategies.

[34] Cave, S., & Ó hÉigeartaigh, S.2018. An AI race for strategic advantage: rhetoric and risks. In AAAI/ACM Conference on Artificial Intelligence, Ethics and Society.