


SCHOOL OF INFORMATION TECHNOLOGY

MASTER OF INFORMATION TECHNOLOGY



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Engineering, Built Environment and
Information Technology

INDIVIDUAL ASSIGNMENT COVER PAGE

Name of Student	Erika Scholtz
Student Number	u11028182
Name of Module	The text arts - NLP approaches to Data Science
Module Code	COS 802
Name of Lecturer	Dr. VN Marivate
Date of Submission	25 November 2021
Contact telephone number	083 648 3878
E-mail address	erika.scholtzz@gmail.com
Declaration:	<i>I declare that this assignment, submitted by me, is my own work and that I have referenced all the sources that I have used.</i>
Signature of Student	
Date received	
Signature of Administrator	
Mark	
Date	
Signature of Lecturer	

COVID-19 Economic Impact Analysis: A Topic Modelling Approach

Erika Scholtz

Student Number: u11028182

Github: [COS-802-Final-Project](#)

erika.scholtzz@gmail.com

ABSTRACT

The global COVID-19 pandemic disrupted the global economy, society, and the way the working world functions. To limit the spread of the virus, governments across the globe implemented drastic lockdown and social distancing measures. Evidence based information is a crucial asset to aid decision and policy making in a time of crisis. This study presents a topic modelling approach to uncover and analyse information about the economic impacts that COVID-19 had on different countries. Separate LDA topic models were created for six different countries to extract the main topics related to the economic impact of COVID-19. The topics obtained for each country were mostly logically coherent and interpretable. Similarities in topics across the different countries were investigated manually and through cosine similarity. It was found that the topics of lockdown restrictions, the number of COVID-19 cases and deaths, the stock market and investment, economic and financial relief measure, and mental health and social wellbeing were present across all country while some other topic such as alcohol and cigarette bans were restricted to one country (South Africa in this example).

Keywords

COVID-19, topic modelling, LDA, NLP, economic impact analysis, machine learning, cosine similarity

1. INTRODUCTION

In December 2019 the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first reported in Wuhan, China. By early 2020 it had spread across the globe and caused the global pandemic now known as COVID-19. This global event has disrupted the global economy, the way in which people live and interact socially, and the way in which the working world functions.

In order to try and limit the spread of the virus, governments and authorities across the globe implemented drastic measures. These measures mostly included isolation and lockdown policies that restricted movement, work, and participation in educational activities [14]. These measures had an immediate and severe impact on the economies of all countries.

Evidence based information is a critical resource to have during a global disaster such as COVID-19. It can be used to inform decision making, form effective healthcare response strategies, and help develop plans for economic reforms. Topic modelling presents a method to extract such useful information from data. It identifies patterns of word occurrences in a set of documents and outputs clusters of words as topics [6]. Topic modelling has been used for a variety of analysis including detecting events from social media data, and classifying trends [6].

This study proposes a topic modelling approach to analyse COVID-19 news data to uncover the economic impacts that the pandemic had on six countries across the world. Sections 2 and 3

present the research questions and problem statement for this study. Section 4 reviews existing literature regarding the economic impact of COVID-19, topic modelling approaches to analysing COVID-19 impacts, and Latent Dirichlet Allocation (LDA) as a topic modelling method. Section 5 explains the research methodology in detail and section 6 presents and discusses the research results. In the conclusion, the study is summarised, limitations are discussed, and future work is proposed.

2. RESEARCH QUESTION

The topic of this research study is “COVID-19 Economic Impact Analysis through a Topic Modelling Approach on News Articles”. This topic is formulated in the field of Computer Science and the area of Machine Learning. The concentration of the study is Natural Language Processing (NLP) with a focus on Topic Modelling.

As a first step in formulating the research question the following question was asked: “Can NLP approaches be applied on news data to analyse the economic impact of COVID-19?”. Two aspects from this question were of particular interest to this study. The first is whether Topic Modelling can be used to highlight the different impacts that COVID-19 had on the economy. The second is how these topics of economic impact might differ across different countries. From these aspects the following research question was derived:

Can topic modelling be used to identify and highlight the different economic impacts that COVID-19 had on countries across the globe?

2.1 Sub-questions

In order to arrive at an answer to the main research question, the two aspects thereof that was mentioned is broken down further into the following sub-questions:

- *Will Latent Dirichlet Allocation (LDA) extract topics that reveal what the economic impacts of COVID-19 were?*
- *Can cosine-similarity be used to reveal the similarities in how groups of countries were affected by COVID-19?*

3. PROBLEM STATEMENT

Early in 2020, COVID-19 caused a devastating global pandemic. Not only have millions of people lost their lives since the virus broke out in December 2019, but the global economic impacts are far reaching. Millions of people across the globe lost their jobs because of the strict lockdown measures implemented in most countries [10]. Increases in trade costs, slow down in production rates, the complete halt of international tourism, and the sudden redirect of activities that require people to be in close proximity all contributed to the global gross domestic product (GDP) declining to 2 percent less than the benchmark [14].

Analysing and knowing both the positive and negative impacts of a global pandemic such as COVID-19 is important in order for governments and authorities to formulate short term and long term response policies [3]. Global cooperation and participation are required to put plans in place to accelerate post-pandemic reform as well as putting effective measures in place to minimise the negative impacts of a potential similar pandemic in the future.

News and media have been valuable sources of information during the COVID-19 pandemic. News outlets across the globe have been reporting on COVID-19 related topics extensively and this makes newspapers a valuable source to understand the social and economic impacts of the pandemic on society [10]. NLP has become prominent in terms of its ability to analyse large amounts of natural language texts. Thus, NLP has been used to analyse the impacts of COVID-19 through methods such as topic modelling.

This study will narrow the focus of the topic modelling approach to extract topics related specifically to economic and economic related activities. It will further build on this NLP approach by comparing the topics extracted across six different countries.

This study is limited to analysing COVID-19 news from the Aylien COVID-19 news dataset. Only one topic modelling algorithm was implemented (LDA). Topic labelling was done manually.

4. LITERATURE SURVEY

Since the outbreak of the global COVID-19 pandemic, research has been focused on investigating and understanding the different aspects and impacts of the pandemic. The impact on the global economy and economic activities has been one of the areas of interest. Natural Language Processing (NLP) approaches to the analysis of COVID-19 based on news articles, social media, scientific studies, and other text based data have been at the forefront of research [10]. In order to form a sufficient background to answering the research questions at hand, it is important to consider five areas. These areas include research on the economic impacts of COVID-19, existing topic modelling approaches to COVID-19 analysis, related topic modelling work on COVID-19 economic impact analysis, LDA, and methods for evaluating topic models.

4.1 Economic Impacts of COVID-19

Because of the complex globalised world of today, the economic impact of a global pandemic is widespread. Understanding what these impacts are, will assist in selecting keywords for this study and evaluating the relevance of topics found.

The first major economic impact of the COVID-19 pandemic was the slow-down of the Chinese economy. This caused an interruption of production and the global supply chain since organisations across the globe are dependent on Chinese production [3]. Limited transport and movement between countries, put pressure on trade and tourism. The start of the pandemic caused panic buying amongst consumers. Stock markets have also plummeted [3]. Maliszewska et al. [14] highlights four economic shocks of COVID-19. The first is a steep decline in employment with the consequence of decreased demand in capital in the market. The second shock was the increase in international trade costs by 25%. The third shock is the sudden decline in tourism which had ripple effects on tourism-related industries such as transportation, accommodation, etc. The last shock is a drastic decline in the demand for services involving close human

contact such as mass transport, and restaurant and recreational activities.

4.2 Topic Modelling Approaches to COVID-19 Analysis

COVID-19 topic modelling approaches have been applied mainly on three types of text data: research papers, news articles, and social media texts.

Alga et al. [1] studied a corpus of COVID-19 research articles to discover what the main topics of research have been. LDA was used as the topic modelling method. Boon-Itt et al. [5] collected and analysed Tweets related to COVID-19 in order to investigate public perception. LDA was used to uncover the main COVID-19 Twitter topics. This was paired with sentiment analysis. It is common in the available COVID-19 NLP literature to combine topic modelling and sentiment analysis in a single study.

Ghasiya et al. [10] presented a topic modelling approach to analysing COVID-19 news topics across four different countries and identified common themes among them. The Top2Vec topic modelling approach was used in this study. Bai et al. [2] analysed COVID-19 news articles through topic evolution and dynamic topic modelling to assist with social management and policy making. The dynamic topic model consisted of a series of LDA models. Liu et al. [13] applied LDA topic modelling in order to investigate health communications and the impact that new media has on the COVID-19 crisis in China. In order to investigate the reactive policies of the Indian government in response to COVID-19, Debnath et al. [7] performed LDA on press information. The topic extraction was approached through considering nudge theory. A keyword co-occurrence network was also constructed to show the network of high frequency words.

4.3 Related Work on Topics Modelling for COVID-19 Economic Impact Analysis

To the knowledge of this author, no other work has been done on COVID-19 data to analyse the economic impact with topic modelling. However, Ghasiya et al. [10] studied general COVID-19 topics across several countries. This study also compared topics found across different countries. In terms of topic modelling being used to analyse the impact of COVID-19, Wright et al. [20] studied the impact of public opinion on public health efforts through topic modelling. A great proportion of past research on COVID-19 topic modelling uses LDA as the topic modelling method.

4.4 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative statistical (probabilistic) model that assumes the text corpus to have a Dirichlet distribution and falls under the area of unsupervised machine learning. Text documents are represented by random mixtures over latent topics [12]. This means that documents are seen as being composed of themes and not just words [15]. LDA was first developed in 2003 by Blei et al. [4]. LDA has been widely applied for topic modelling across a wide spectrum of fields including political science, medical science, social media analysis, software engineering, etc. LDA is still one of the most popular topic modelling methods [12].

4.5 Measures for Evaluating Topic Models

In order to evaluate a topic model, there are three main methods that are generally used. These three methods include, human

evaluation, coherence, and perplexity [6]. The methods can be summarized as follows:

1. **Human evaluation:** This method involves human judgement. It takes into consideration to what extent a human evaluator would agree with and make sense of the topic assignment by the algorithm. This method might introduce some bias into topic evaluation.
2. **Coherence:** This method determines how closely related the words within a topic's top N words are [16]. Thus, a good topic will be coherent.
3. **Perplexity:** This is a measure of how accurately a trained model will be able to correctly assign an unseen document to a topic from the model [6].

Coherence and human evaluation are thus the methods used to determine how interpretable topics are.

5. METHODOLOGY

Figure 1 shows the design of the methodology followed in this study. The first step is to filter the dataset down to selected countries and keywords. The data is then pre-processed to clean and prepare the texts for analysis. Thereafter bigram and trigram modelling, as well as TF-IDF is performed on the dataset to prepare it for topic modelling. A separate LDA topic model is created for each country. The results of the topic models are evaluated and compared as a last step. The sections to follow explain each step in more detail.

5.1 Dataset

The dataset that was used for this study is the Aylien Coronavirus News Dataset that is free to use for non-commercial projects [9]. The dataset consists of 1.67M news articles from different countries that is related to the COVID-19 pandemic. The dataset includes English news articles from November 2019 to July 2020.

5.1.1 Selected Countries

A subset of six different countries was taken from the full dataset. The countries include the United States (US), United Kingdom (GB), India (IN), Canada (CA), Australia (AU), and South Africa (ZA). The dataset was filtered to include only these countries.

5.1.2 Keyword Selection

In order to achieve the economic impact analysis of this study, the dataset for the six countries was filtered to include only articles that contain certain economy related keywords. These keywords were formulated by the literature study and trial and error. The keywords include the following: “business”, “businesses”,

“economy”, “economic”, “GDP”, “gross domestic product”, “jobs”, “unemployment”, “industry”, “trade”, “shortage”, “panic buying”, “stocks”, “stock”, “market”, “investment”.

A final set of 10,000 news articles was then randomly selected from the filtered dataset for each country with the exception of South Africa that only has 1,843 articles available. Table 1 summarises the final dataset used in this study.

5.2 Pre-processing

In order to ensure that the best possible results are obtained from the LDA topic models, the text data was first pre-processed to clean and prepare it for topic modelling. Pre-processing can have profound effects on the results of unsupervised NLP algorithms [8]. The Aylien Coronavirus News Dataset is already processed to an extent. Further pre-processing involved the following: filtering to include only words from certain parts of speech, lemmatisation, removing stopwords, lowercase and deaccent, and tokenisation. Pre-processing was done using Python's Gensim and spaCy libraries.

5.2.1 Part of Speech Filtering and Lemmatisation

For these pre-processing steps, a spaCy English language pipeline that is optimised for web and news texts was loaded and used. The dataset was split into individual tokens. These tokens were then filtered to include only nouns, adjectives, verbs, and adverbs. This removes any special characters and punctuations from the texts. The tokens/words were then lemmatised. Lemmatisation is the process of converting words to their root form [6]. This can simplify the text corpus and improve topic modelling results.

Default English stopwords from the spaCy language pipeline was loaded and used to filter out stopwords from the dataset. Additional stopwords were added to this list that were influencing the topic models negatively. Stopwords were removed before and after lemmatisation.

5.2.2 Lowercasing, De-accentuating, and Tokenisation

The lemmatised texts were further pre-processed by lowercasing each word and removing accents on letters by using Gensim's `simple_preprocess` function. The last step in pre-processing the data was to tokenise the texts.

5.3 Bigram and Trigram Modelling

Studies have shown that incorporating n-grams in topic modelling approaches can improve results [18]. Bigrams and trigrams are two and three words frequently occurring together in a text document. An example in the COVID-19 context would be “social distancing”. The tokenised text was converted to include frequent bigrams and trigrams from the articles to improve topic modelling results.

5.4 TF-IDF

In this step TF-IDF was used to remove words that occur below a

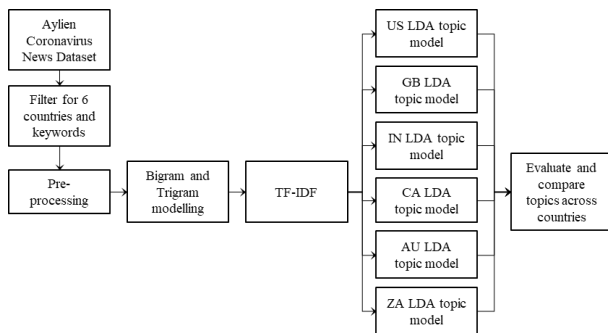


Figure 1: Research methodology

Table 1: Number of news articles for each country

Country	Number of news articles
United States (US)	10,000
United Kingdom (UK)	10,000
India (IN)	10,000
Canada (CA)	10,000
Australia (AU)	10,000
South Africa (ZA)	1,834

specified threshold frequency. Also done in this step was to remove words that occur across every news articles. This removed words that would not necessarily contribute to distinguishing topics from the text input. The remaining words were then used to create the bag of words corpus and dictionary that serves as input to the LDA topic models.

5.5 LDA Topic Modelling

The main research question of this study is whether topic modelling can successfully be used to identify the economic impacts that COVID-19 had on different countries. In order to do this, a topic modelling approach is required. LDA is used as the topic modelling algorithm in this study. A separate LDA model was created for each of the countries being studied in order to be able to compare the topics across different countries.

For each of the LDA models the model parameters as well as the number of topics chosen were optimised in an iterative approach. The Python Gensim library was used to create the LDA models. The Ldamulticore model was used for its ability to process the models using multiple cores in parallel.

5.5.1 LDA Parameter Optimisation

The following model parameters were tuned for each country's LDA model:

- The Alpha parameter (topic-document density) was chosen as asymmetric.
- They chunk size was found to be optimal for coherence as well as computational performance when set to 500. The chunk size represents the number of documents that the algorithm will consider at once.
- The number of passes was set to 20. This is the number of times the algorithm passes over the whole corpus.

5.5.2 Number of Topics

The optimal number of topics for each LDA model was found by plotting the coherence score against the number of topics for a range of topic numbers. Between 10 and 36 topics were considered. The marginal topic distribution and topic overlap were also considered when selecting the number of topics. If the topic distributions were too small or if there were much topic overlap, the number of topics were lowered.

5.5.3 Topic Visualisation and Labelling

As a last step in the topic modelling process for each country, the topic models were visualised by using the Python pyLDAvis library. The visualisations include an intertopic distance map for each model that also shows the marginal topic distributions. It also includes a visual showing the top 20 most salient terms for each topic. These visualisations were used to aid in the manual labelling of each of the topics in each of the five models. Topic labelling was done based on the top 20 words in each topic.

5.6 Model Evaluation and Comparison

Model evaluation and comparison consist mainly of two parts. The first is to evaluate how well LDA performed in extracting topics related to COVID-19 economic impact for each country. The second part is to determine if there are any similarities in the topics found for each of the five countries.

5.6.1 Topic Evaluation for each country

The LDA models for each country were evaluated using coherence, and human judgement and interpretability [6]. The c_v coherence measure was used. C_v coherence works on a sliding window. It segments the top words using one-set method and cosine similarity and normalized pointwise mutual information (NPMI). Human judgement and visual inspection were performed to ensure that there is minimal overlap between topics and that sense can be made of the topics. Since the focus of this study was to extract valuable information from the topic models and not to predict on unseen data, perplexity was not used.

5.6.2 Topic Similarity

Two methods were used to compare the topics found between the different countries. The first was human evaluation of similarities. The second method was to calculate the cosine similarity between the top 20 words in each topic of a country and every other topic in all other countries. The results were plotted as heatmaps between the different country pairs. From this the topics with the highest similarity scores from each country pair was found.

6. EXPERIMENTAL RESULTS AND DISCUSSION

The purpose of the experiments performed was to answer the research questions posed in section 2. Thus, six LDA models were created and evaluated (one for each country). These models were then compared to find any similarities amongst the topics found for the different countries.

6.1 Bigrams, Trigrams, and TF-IDF

After the dataset was filtered and pre-processed, a dictionary and corpus (bag of words) were created for each of the 6 countries under question. These dictionaries and corpuses were used as input to train and obtain 6 LDA topics models.

It was found that including bigrams and trigrams in the corpus improved the coherence scores of the LDA models for each country. By using TF-IDF to filter out words below the threshold TF-IDF value of 0.03 and remove words present in all articles, the coherence scores were further improved.

6.2 Optimal Number of Topics Found

As mentioned in the methodology, the coherence scores for different numbers of topics were evaluated for each country. An example of the results is shown in Figure 2. It can be seen that the number of topics for the US LDA model that results in the highest coherence score is 14. The figures showing the number of topics vs. coherence scores obtained for the other five countries can be seen in Appendix A. Table 2 summarises the coherence scores

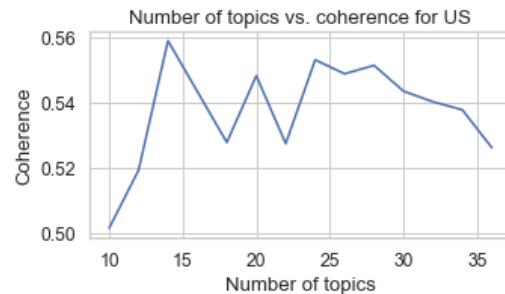


Figure 2: Number of topics vs coherence for US

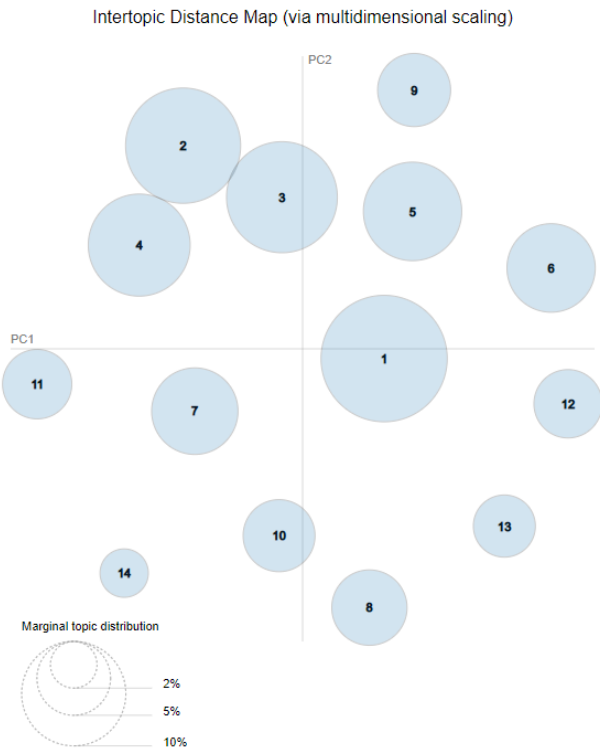
Table 2: Number of topics optimisation results

k	US	GB	IN	CA	AU	ZA
10	0.502	0.511	0.514	0.538	0.505	0.437
12	0.519	0.540	0.533	0.528	0.529	0.411
14	0.559	0.529	0.538	0.528	0.519	0.447
16	0.543	0.528	0.543	0.552	0.500	0.464
18	0.528	0.548	0.546	0.534	0.523	0.456
20	0.548	0.535	0.560	0.537	0.539	0.478
22	0.527	0.562	0.559	0.542	0.541	0.489
24	0.553	0.553	0.567	0.552	0.540	0.474
26	0.549	0.539	0.559	0.565	0.524	0.482
28	0.551	0.544	0.540	0.537	0.540	0.449
30	0.544	0.537	0.538	0.545	0.527	0.453
32	0.540	0.547	0.548	0.541	0.534	0.433
34	0.538	0.541	0.513	0.533	0.530	0.459
36	0.526	0.542	0.538	0.551	0.534	0.471

obtained for each country at different numbers of topics (k). The highest coherence scored obtained for each is highlighted.

The LDA models obtained at the highlighted number of topics for each country were visualised with pyLDAvis to further evaluate the optimal number of topics. Figure 3 shows the intertopic distance map obtained for the US LDA model. This plot represents the Jensen-Shannon divergence between the topics that is plotted in 2D by using multidimensional scaling. The size of each circle represents what percentage of the total corpus that topic makes up. This plot was used to further improve on the topics obtained for each country by looking at two things:

- Overlap between topics. If there was too much overlap, the number of topics was decreased.
- Marginal topic distribution. If there were too many non-

**Figure 3: Intertopic distance map for the United States**

prevalent topics present, the number of topics was decreased.

The above steps were repeated until a satisfactory intertopic distance map was achieved for each country. The intertopic distance maps for the other five countries can be seen in Appendix B. Table 3 summarises the final number of topics used for each country's topic model as well as the coherence scores obtained for the final models. These results are discussed in section 6.4.

6.3 Topic Labelling

The top 20 words obtained from the LDA models for each country was used to manually label each topic for each country. The interactive pyLDAvis visuals in Python were also used to assist with the labelling process. Labelling was performed before cosine similarity was done on the topics to avoid confirmation bias.

Table 4 to Table 9 show the top 5 most prevalent topics for each country. Also shown are some of the top 20 words of each topic that were used to formulate the topic labels. The top 5 topics for the US are around mental health, lockdown restrictions on stores and restaurants, the stock market and investment, law and policy, and COVID-19 testing and vaccines. The top 5 topics for the United Kingdom are public health response, sports and entertainment, financial relief and loans, the stock market and investment, and lockdown restrictions on pubs and restaurants. The top 5 topics for India include economic support measures, economic growth, COVID-19 cases and deaths, lockdown restrictions of stores, and social wellbeing. The top 5 topics found for Canada include social wellbeing, lockdown restrictions, COVID-19 cases and spread, government economic support measures, and healthcare and healthcare workers. The top 5 topics for Australia were found to be lockdown restrictions, COVID-19 cases and deaths, social wellbeing, government economic plans, and economic and market growth. Lastly, the top 5 topics for South Africa include government COVID-19 response, public healthcare, COVID-19 cases and deaths, social wellbeing, and government relief programs.

6.4 Topic Model Performance Evaluation and Discussion

6.4.1 Coherence Scores

As mentioned, Table 3 shows the coherence scores obtained for each of the countries' LDA models. The best coherence score of 0.560 was obtained for India with 20 topics, followed closely by the US LDA model with a coherence score of 0.559 with 14 topics. The worst coherence score of 0.48 was obtained for South Africa, which is in line with the expectation that a topic model would perform worse on a smaller set of articles.

6.4.2 Human Interpretability

The performance of the topics was not only evaluated by the coherence scores, but also by how easy they are to interpret by a

Table 3: Final LDA models number of topics and coherence scores

Country	Number of topics	Coherence
US	14	0.559
GB	18	0.548
IN	20	0.560
CA	16	0.552
AU	22	0.541
ZA	22	0.489

Table 5: Top 5 topics for the United States

Topic	Top words	Topic assigned
1	time, need, feel, people, help, crisis, talk, tell, change...	Mental health
2	reopen, business, open, store, restaurant, close, plan...	Lockdown restrictions of stores and restaurants
3	stock, investor, market, share, price, economy, fall, demand...	Stock market and investment
4	government, response, police, lawmaker, president, bill...	Law and policy
5	test, people, report, death, vaccine, outbreak, case...	Covid testing and vaccines

Table 6: Top 5 topics for the United Kingdom

Topic	Top words	Topic assigned
1	government, pandemic, decision, plan, warn, country...	Public health response
2	club, player, season, game, fan, film, star, sport, team...	Sports and entertainment
3	support, government, loan, financial, help, fun, scheme...	Financial relief and loans
4	stock, price, market, trade, investor, oil, fall, demand...	Stock market and investment
5	close, reopen, open, pub, restaurant, lockdown, rule...	Lockdown restrictions on restaurants and pubs

Table 4: Top 5 topics for India

Topic	Top words	Topic assigned
1	country, support, crisis, economy, plan, measure...	Economic support measures
2	economy, growth, impact, decline, revenue, fall, low...	Economic growth
3	case, test, people, death, infection, virus, report, number	Covid cases and deaths
4	police, shop, open, lockdown, restriction, essential, allow...	Lockdown restrictions on shops
5	people, health, social, change, community, problem, need...	Social wellbeing

human. The topics obtained for the US were logically coherent with minimal overlap. For the United Kingdom, India, Canada, and Australia the topics were still logically coherent, but it was slightly more difficult to label the topics. There seems to be slight overlap in some topics for these countries. As an example, Canada presented two topics that both seem to be related to the spread, testing, number of cases and number of deaths of the coronavirus. The South Africa LDA model produced topics that were the most difficult to interpret; however, most topics were still logically coherent.

Even though the topics found by the LDA models were coherent, it is difficult to interpret what the relation is between the topics and the economic impact. As an example, some topics relate to the stock market and investment, but does not contain information as to whether the market and investments are up or down. The COVID-19 cases and testing topic is also very prominent but does not necessarily give much information about the economic impact.

6.5 Cross Country Topic Comparison

6.5.1 Manual Comparison

From the topic results for each country there are some topic themes that are present across all six countries. The themes that all countries have in common are firstly lockdown restrictions that

Table 7: Top 5 topics for Canada

Topic	Top words	Topic assigned
1	need, family, community, right, home, people, want, time	Social wellbeing
2	reopen, open, public, allow, restriction, physical distancing	Lockdown restrictions
3	outbreak, virus, spread, case, test, infection, symptom...	Covid cases and spread of the virus
4	support, program, pandemic, federal, funding, financial...	Government economic support measures
5	health, worker, supply, patient, hospital, shortage, doctor, care..	Healthcare and healthcare workers

Table 8: Top 5 topics for Australia

Topic	Top words	Topic assigned
1	open, reopen, restriction, border, lockdown, measure...	Lockdown restriction
2	case, test, infection, death, outbreak, number, spread...	Covid cases and deaths
3	family, want, feel, people, work, home, good, life, try...	Social wellbeing
4	economic, government, plan, crisis, future, opportunity...	Government economic plan
5	market, economy, share, impact, growth, drop, low...	Economic and market growth

Table 9: Top 5 topics for South Africa

Topic	Top words	Topic assigned
1	economy, crisis, support, government, business, country..	Government covid response
2	Hospital, patient, work, staff, virus, province, department...	Public healthcare
3	Case, death, infection, report, people, record, spread, rise...	Covid cases and deaths
4	Community, help, family, child, hunger, vulnerable...	Social wellbeing
5	Taxi, grant, fund, receive, government, regulation...	Government relief programs

were imposed on the public and businesses. This could be one of the main contributors of the fourth economic shock mentioned by Maliszewska et al. [14] which is the drastic decline in demand for services involving close contact. The second theme present across all countries is topics that talk to the stock market, investments, and economic growth. Baldwin et al. [3] mentions stock markets plummeting as one of the impacts of COVID-19. Across all six countries there are topics present that point to financial and economic support programs and measures provided by government. This points the to need to provide relief to businesses and individuals who are feeling the financial pressures related to the COVID-19 economic shocks.

An interesting theme present across all countries is that of social wellbeing, family care, and mental health. The pandemic did not only have physical medical impacts, but also caused stress and trauma in the population. The United States and Canada has topics more specific to mental health whereas the other country focused more on the social aspects in general.

Education was also a topic present for all countries. Hunashek et al. [11] highlights how the learning losses caused by COVID-19 will follow student into the job market and have negative economic outcomes in the future. All countries spoke about the covid cases and deaths, and some countries also mentioned the vaccine. The vaccine brings hope to the world to return back to

normal and reinstate economic stability. All countries spoke about sports and entertainment except for South Africa.

There is a theme of protests present in the United States, United Kingdom, Canada, and Australia. Some of the top words in these topics point to it possibly being about the Black Lives Matter movement. An interesting theme was present in the United States, India, and Canada and this is around digital solutions and technology. This is one of the positive impacts of Covid-19 and saw tech giants such as Microsoft stronger than even before [19]. Australia is the only country where panic buying came out as a topic. South Africa was differentiated by topics of corruption, alcohol and cigarette bans, and churches and places of worship. Canada mentioned the construction and infrastructure industry. Australia is the only country mentioning something about the property market. This can point to the crisis in the housing market brought about by COVID-19 [17].

6.5.2 Cosine Similarity of Topics

One of the research questions of this study is whether cosine similarity between topics of the different countries can be used to determine the similarities in the economic impacts of Covid-19 on those countries.

As mentioned in the methodology, the cosine similarity between the different countries' topics were calculated by using the top 20 words of each topic. Figure 4 shows an example of one of the cosine similarities between the topics obtained for the United States and the United Kingdom. The darker green areas point to topics that are the most similar between the two countries. The cosine similarity heatmaps for the other country pairs can be found in Appendix C. These heatmaps were used to identify the top three most similar topics between the different country pairs. Table 4 summarises the most similar topics with their cosine similarity scores for each country pair. The country pairs with the highest similarities in topics are the United States vs. the United Kingdom, the United States vs. Australia, and the United States vs. Canada. This is in line with the fact that these countries have similar economies. The country pair with the lowest topic similarity scores are Canada vs. South Africa. The topic similarities presented in Table 4 are in line with the similarities

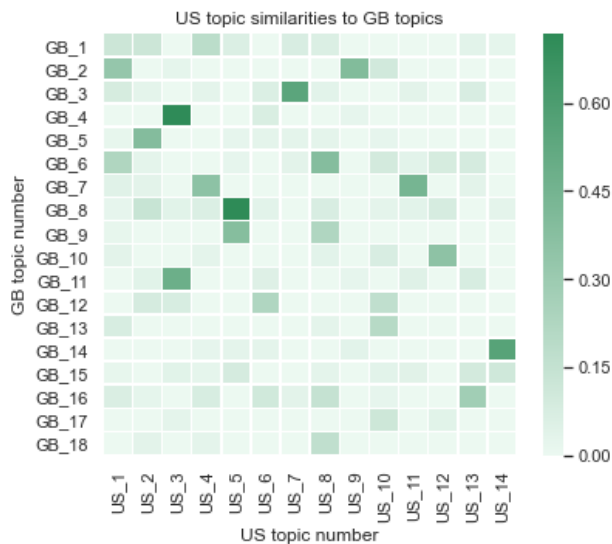


Figure 4: Topic cosine similarity between the US and GB

uncovered through manual inspection.

7. CONCLUSION AND FUTURE WORK

There is no doubt that COVID-19 brought about sudden and significant changes to the world and society. Despite drastic public health efforts and promising vaccine rollouts, COVID-19 will continue to challenge the world and its economies. Understanding these economic impacts will aid the formulation of public and private response efforts and help focus resources where it is most needed.

This study presented a topic modelling approach to analyse the economic impacts of COVID-19 across six different countries. The Aylien Coronavirus News dataset was filtered for economy related articles and used as input to the topic modelling. A corpus of 10,000 articles was used for each topic model for the United States, United Kingdom, India, Canada, and Australia. A smaller

Table 4: Top 3 most similar topics for different country pairs

Country pair	Topics	Cosine similarity
US vs GB	Stock market and investment	0.714
	Covid cases and deaths and testing	0.719
	Travel and tourism	0.563
US vs IN	Online business and digital tech	0.708
	Covid cases and deaths	0.572
	Healthcare and healthcare workers	0.456
US vs CA	Healthcare and healthcare workers	0.730
	Mental and social wellbeing	0.675
	Covid cases, testing and deaths	0.601
US vs AU	Covid cases, testing and deaths	0.773
	Healthcare and healthcare workers	0.700
	Stock markets and investment	0.601
US vs ZA	Covid cases and deaths	0.547
	Politics and elections	0.446
	Mental and social wellbeing	0.428
GB vs IN	Covid cases and deaths	0.566
	Economic growth	0.463
	Stock market and investment	0.436
GB vs CA	Covid cases and testing,	0.547
	Stock market and investing	0.514
	Politics and elections	0.421
GB vs AU	Covid cases and deaths	0.745
	Economic and market growth	0.622
	Retail sales and panic buying	0.500
GB vs ZA	Coronavirus cases and deaths	0.587
	Politics and elections	0.498
	Healthcare, covid testing & vaccine	0.428
IN vs CA	Covid cases and deaths	0.636
	Healthcare and healthcare workers	0.488
	Markets, trade and economy	0.485
IN vs AU	Mental health and social wellbeing	0.601
	Covid cases and deaths	0.582
	Economic and market growth	0.536
IN vs ZA	Social wellbeing	0.584
	Covid cases and deaths	0.483
	Education	0.364
CA vs AU	Healthcare and healthcare workers	0.619
	Covid cases and deaths	0.611
	Social wellbeing	0.607
CA vs ZA	Covid cases and testing	0.456
	Social wellbeing	0.395
	Lockdown restrictions	0.376
AU vs ZA	Social wellbeing,	0.751
	Covid cases and deaths	0.572
	Healthcare and healthcare workers	0.472

corpus of 1,834 articles was available for the South Africa model. A LDA topic model was trained for each of the six countries to uncover the topics of economic impact most prevalent in each country. To this author's knowledge this was the first study to analyse the economic impact of COVID-19 through a topic modelling approach.

The LDA models for each country extracted mostly logically coherent topics for each country. The coherence scores obtained for the models were satisfactory for the United States, United Kingdom, India, Canada, and Australia and ranged between 0.541 and 0.560. The coherence score obtained for South Africa was significantly lower than that of the other countries at 0.489. It is suggested that a larger dataset is obtained for South Africa to rerun the experiments on. Human interpretation and the cosine similarity method used uncovered similarities in the topics found across the different countries. It was uncovered that the topics of lockdown restrictions, the number of COVID-19 cases and deaths, the stock market and investment, economic and financial relief measure, and mental health and social wellbeing were present across all countries. Some countries presented topic that were unique to that country, e.g., South Africa was the only country with a topic about alcohol and cigarette bans. Education was also highlighted as one of the topics that will carry significant economic impacts in the labour market into the future. One positive economic topic that was extracted for the United States, India, and Canada, was the boom in the digital technologies market due to the rapid digital transformation that the COVID-19 lockdown and social distancing measures brought about.

This study was limited using news articles collected by the Aylien News API. As mentioned, there was also a limited number of articles available for South Africa. It is suggested that other news sources should be included in future experiments to expand the corpus on which the models are trained. The experiments were limited to the use of LDA as the topic modelling method. Even though the topics were relatively coherent, it was difficult to see exactly how the topics relates to economic impacts. Future work would include expanding these experiments to include and compare it to other topic modelling approached such as Latent Semantic Allocation (LSA), and Non-Negative Matrix Factorisation (NMF). The study should also be expanded and run on more countries from different types of economies across the globe. To further improve the interpretability of the topics found, it is suggested that this study is combined with sentiment analysis. This would help to determine is the economic impact topics that were uncovered point to negative or positive economic impacts. This study could also be expanded to topic modelling and sentiment analysis on social media texts. This can highlight the economic impacts from a public point of view.

8. REFERENCES

- [1] Älgå, A., Eriksson, O. and Nordberg, M. 2020. Analysis of Scientific Publications During the Early Phase of the COVID-19 Pandemic: Topic Modeling Study. *Journal of Medical Internet Research*. 22, 11 (Nov. 2020), e21559. DOI:<https://doi.org/10.2196/21559>.
- [2] Bai, Y., Jia, S. and Chen, L. 2020. Topic Evolution Analysis of COVID-19 News Articles. *Journal of Physics: Conference Series*. 1601, (Aug. 2020), 052009. DOI:<https://doi.org/10.1088/1742-6596/1601/5/052009>.
- [3] Baldwin, R.E., Weder, B., and Centre for Economic Policy Research (Great Britain) 2020. Economics in the time of COVID-19. CEPR Press.
- [4] Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3, Jan (2003), 993–1022.
- [5] Boon-Itt, S. and Skunkan, Y. 2020. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health and Surveillance*. 6, 4 (Nov. 2020), e21978. DOI:<https://doi.org/10.2196/21978>.
- [6] Culmer, K. and Uhlmann, J. 2021. Examining LDA2Vec and Tweet Pooling for Topic Modeling on Twitter Data. *WSEAS TRANSACTIONS ON INFORMATION SCIENCE AND APPLICATIONS*. 18, (Jul. 2021), 102–115. DOI:<https://doi.org/10.37394/23209.2021.18.13>.
- [7] Debnath, R. and Bardhan, R. 2020. India nudges to contain COVID-19 pandemic: A reactive public policy analysis using machine-learning based topic modelling. *PLoS ONE*. 15, 9 (Sep. 2020), 1–25. DOI:<https://doi.org/10.1371/journal.pone.0238972>.
- [8] Denny, M.J. and Spirling, A. 2018. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*. 26, 2 (Apr. 2018), 168–189. DOI:<http://dx.doi.org/10.1017/pan.2017.44>.
- [9] Free Coronavirus News Dataset – Updated: <https://aylien.com/blog/free-coronavirus-news-dataset>. Accessed: 2021-11-16.
- [10] Ghasiya, P. and Okamura, K. 2021. Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach. *IEEE Access*. 9, (2021), 36645–36656. DOI:<https://doi.org/10.1109/ACCESS.2021.3062875>.
- [11] Hanushek, E.A. and Woessmann, L. The Economic Impacts of Learning Losses. 24.
- [12] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*. 78, 11 (Jun. 2019), 15169–15211. DOI:<https://doi.org/10.1007/s11042-018-6894-4>.
- [13] Liu, Q., Zheng, Z., Zheng, J., Chen, Q., Liu, G., Chen, S., Chu, B., Zhu, H., Akinwunmi, B., Huang, J., Zhang, C. and Ming, W.-K. 2020. Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: A Digital Topic Modeling Approach (Preprint). *Journal of Medical Internet Research*. 22, (Apr. 2020). DOI:<https://doi.org/10.2196/19118>.
- [14] Maliszewska, M., Mattoo, A. and van der Mensbrugghe, D. 2020. *The Potential Impact of COVID-19 on GDP and Trade: A Preliminary Assessment*. World Bank, Washington, DC.
- [15] Mifrah, S. and Benlahmar, E.H. 2020. Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID-19 Corpus. *International Journal of Advanced Trends in Computer Science and Engineering*. (Aug. 2020). DOI:<https://doi.org/10.30534/ijatcse/2020/231942020>.
- [16] Syed, S. and Spruit, M. 2017. Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. 2017 *IEEE International Conference on Data*

Science and Advanced Analytics (DSAA) (Oct. 2017), 165–174.

- [17] The Covid-19 Pandemic Has Fueled A Crisis In The Housing Market: <https://www.forbes.com/sites/saibala/2021/04/27/the-covid-19-pandemic-has-fueled-a-crisis-in-the-housing-market/>. Accessed: 2021-11-20.
- [18] Using phrases and document metadata to improve topic modeling of clinical reports | Elsevier Enhanced Reader: <https://reader.elsevier.com/reader/sd/pii/S1532046416300284?token=E1258DE07B782C3FDC2E8BE57D28CB477600BC6D26995280136AE298873D773830841ED06449B1B615E124035AFA256B&originRegion=eu-west-1&originCreation=20211116103859>. Accessed: 2021-11-16.
- [19] Wakabayashi, D., Nicas, J., Lohr, S. and Isaac, M. 2020. Big Tech Could Emerge From Coronavirus Crisis Stronger Than Ever. *The New York Times*.
- [20] Wright, L., Burton, A., McKinlay, A., Steptoe, A. and Fancourt, D. 2021. *Public Opinion about the UK Government during COVID-19 and Implications for Public Health: A Topic Modelling Analysis of Open-Ended Survey Response Data*. Public and Global Health.

About the authors:

Erika Scholtz is currently pursuing a Master's in Information Technology is the field of Big Data Science at the University of Pretoria. She is an experienced Data Engineer and BI Specialist with a focus on the financial industry. She also has experience in business strategy consulting and mechanical project engineering. Erika obtained her undergraduate degree in Mechanical Engineering at the University of Pretoria in 2015.

9. APPENDIX A

Figure 5 to Figure 9 show the results obtained for number of topic optimisation for the LDA models for each of the six countries. It plots the coherence score against the number of topics.

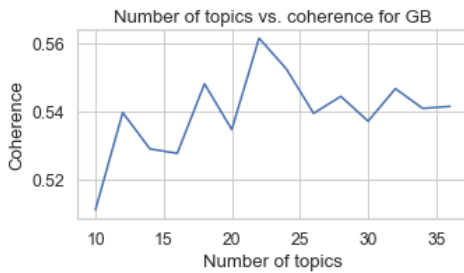


Figure 5: Number of topics vs. coherence for GB

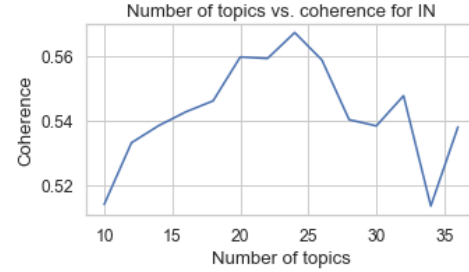


Figure 6: Number of topics vs. coherence for IN

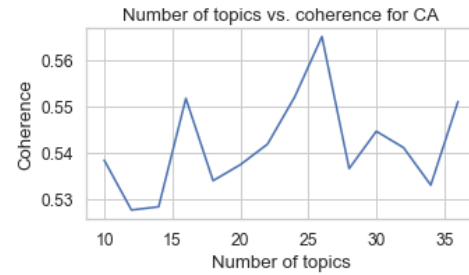


Figure 7: Number of topics vs. coherence for CA

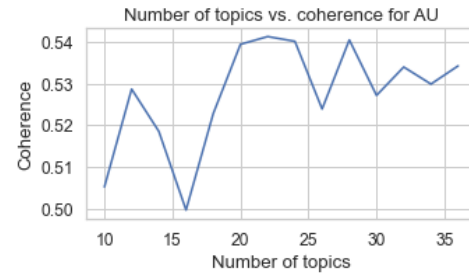


Figure 8: Number of topics vs coherence for AU

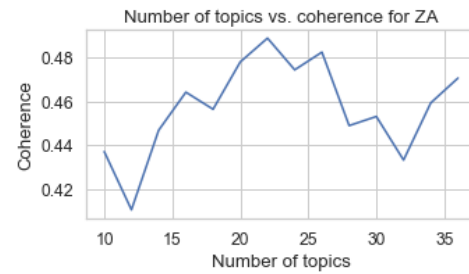


Figure 9: Number of topics vs. coherence for ZA

10. APPENDIX B

Figure 10 to Figure 14 show the pyLDavis intertopic distance maps obtained for the LDA models for each country.

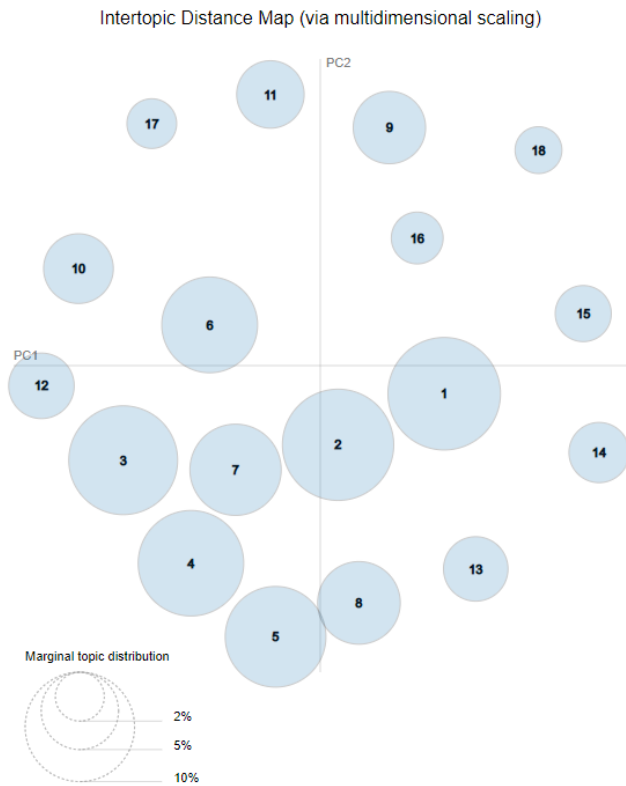


Figure 10: Intertopic distance map for GB

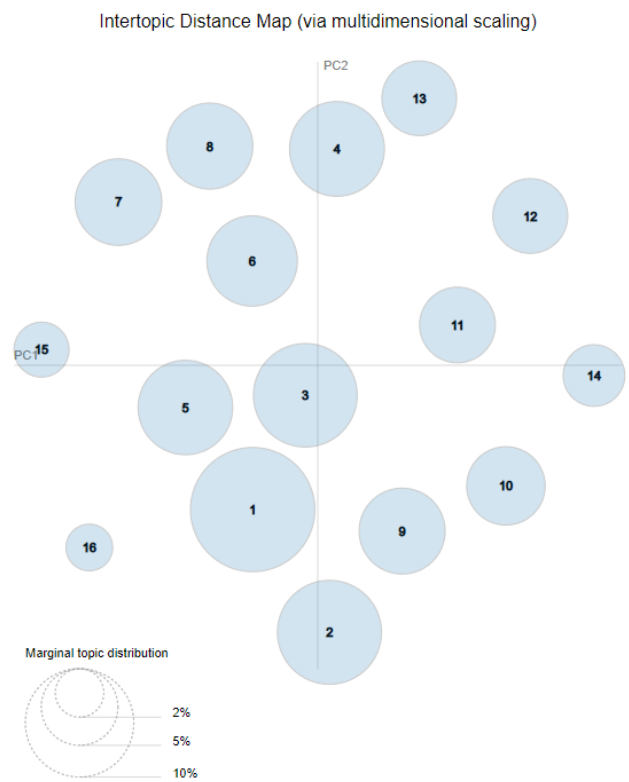


Figure 12: Intertopic distance map for CA

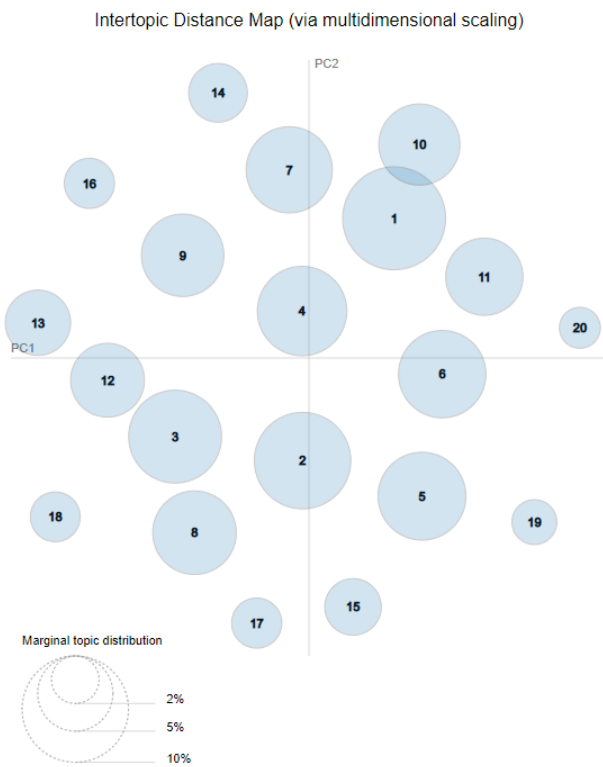


Figure 11: Intertopic distance map for IN

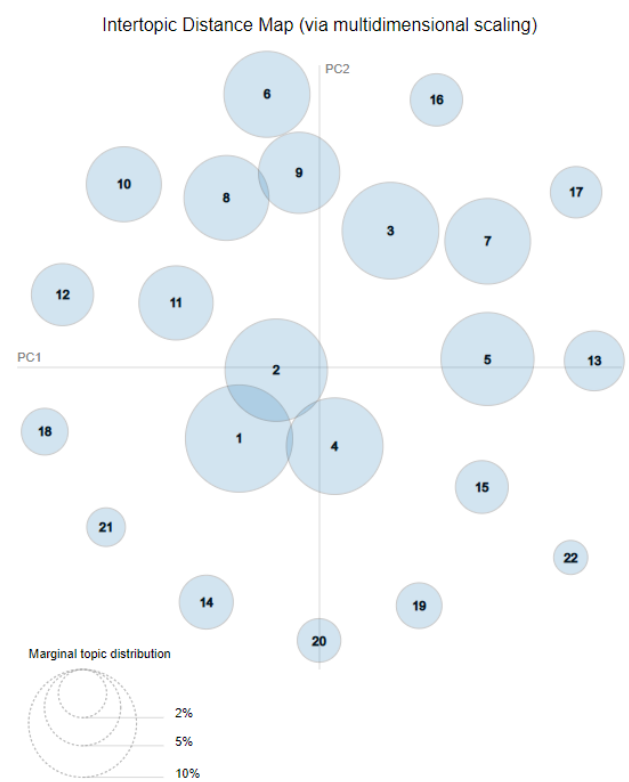


Figure 13: Intertopic distance map for AU

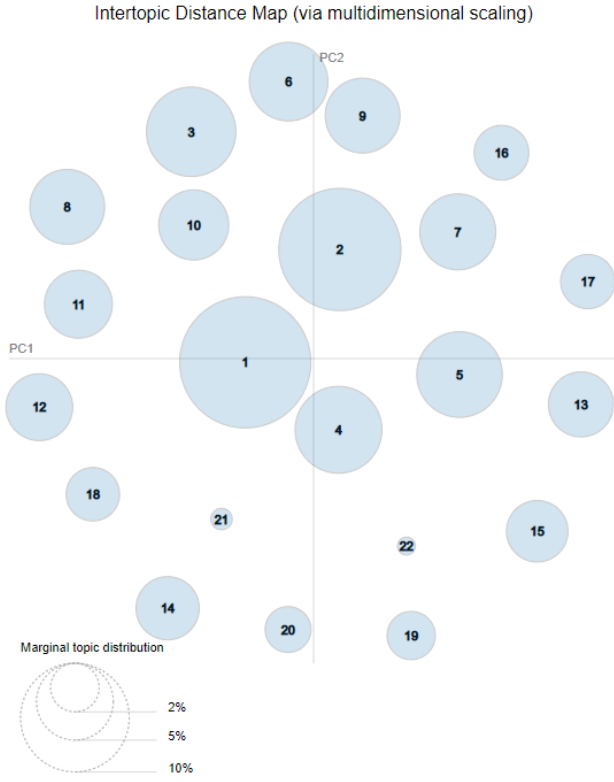


Figure 14: Intertopic distance map for ZA

11. APPENDIX C

Figure 15 to Figure 28 shows the heatmaps of the cosine similarities between topics of different country pairs.

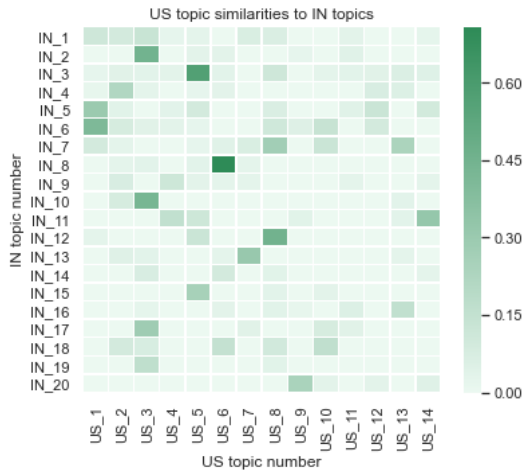


Figure 15: Topic cosine similarity between the US and India

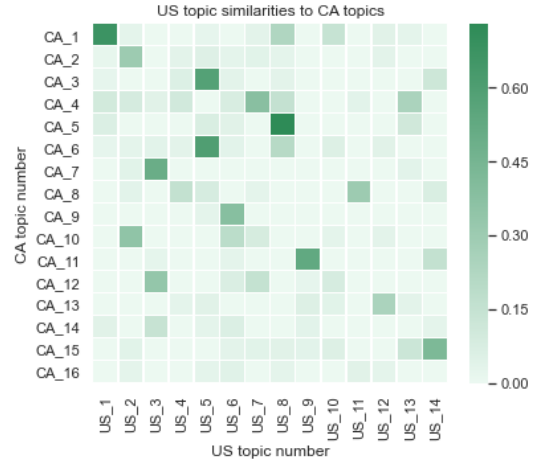


Figure 16: Topic cosine similarity between the US and Canada

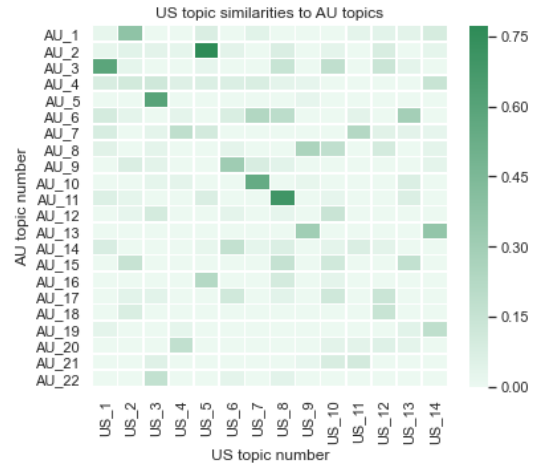


Figure 17: Topic cosine similarity between the US and AU

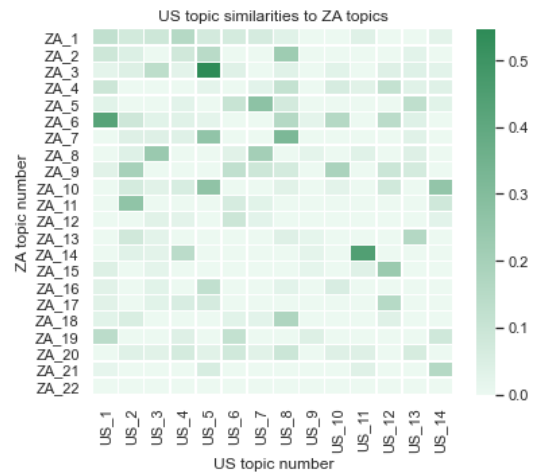


Figure 18: Topic cosine similarity between the US and ZA

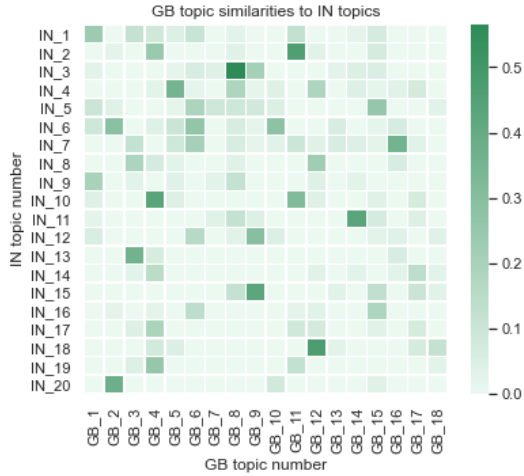


Figure 19: Topic cosine similarity between GB and India

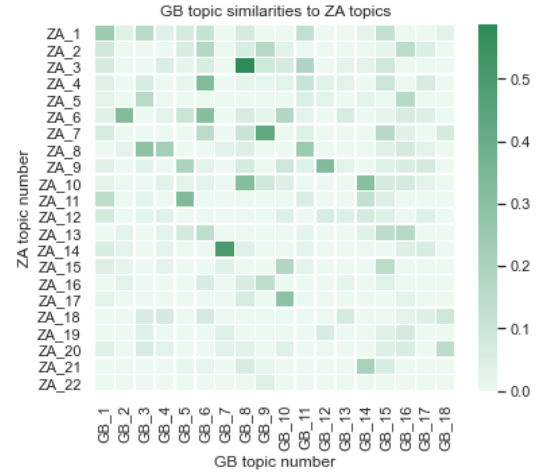


Figure 22: Topic cosine similarity between GB and ZA

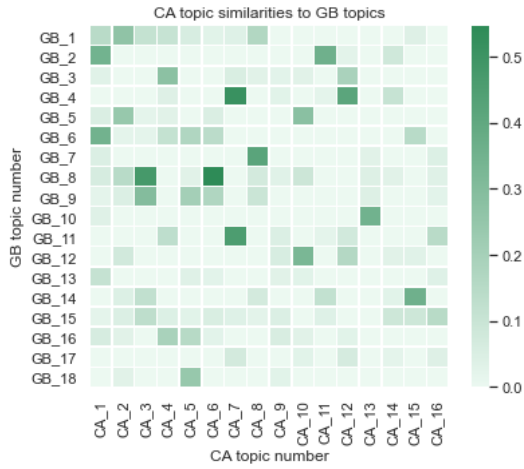


Figure 20: Topic cosine similarity between CA and GB

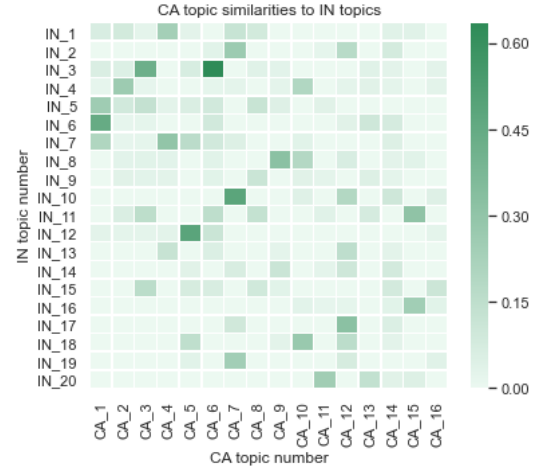


Figure 23: Topic cosine similarity between CA and IN

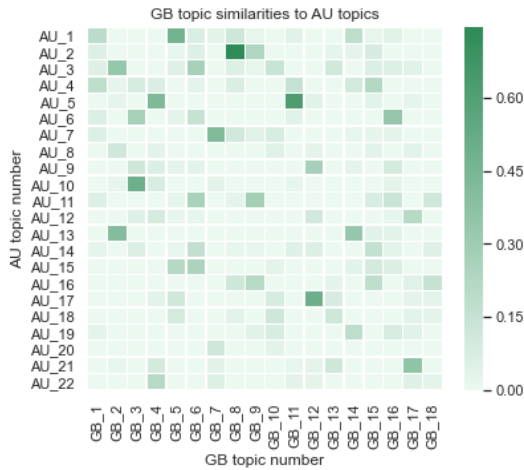


Figure 21: Topic cosine similarity between GB and AU

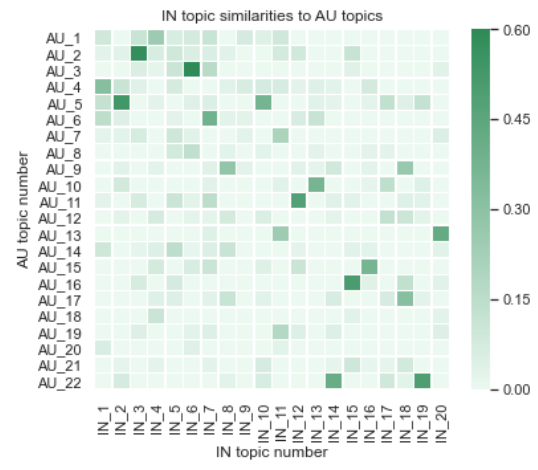


Figure 24: Topic cosine similarity between IN and AU

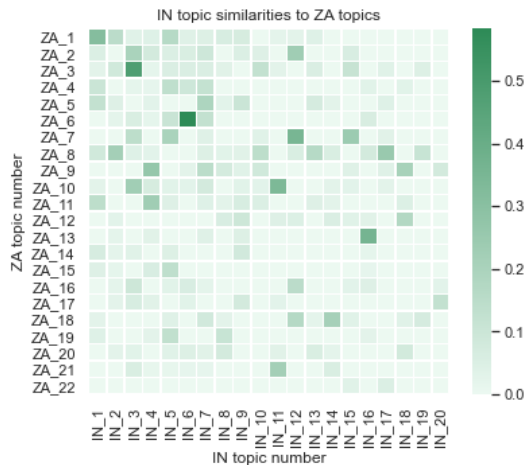


Figure 25: Topic cosine similarity between IN and ZA

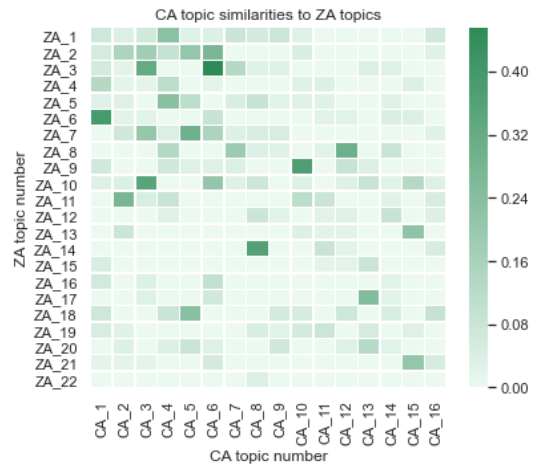


Figure 27: Topic cosine similarity between CA and ZA

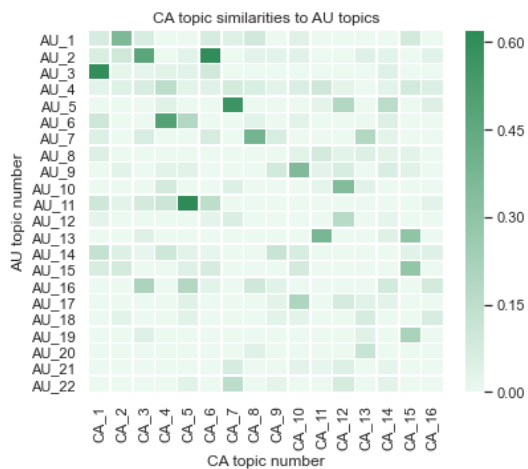


Figure 26: Topic cosine similarity between CA and AU

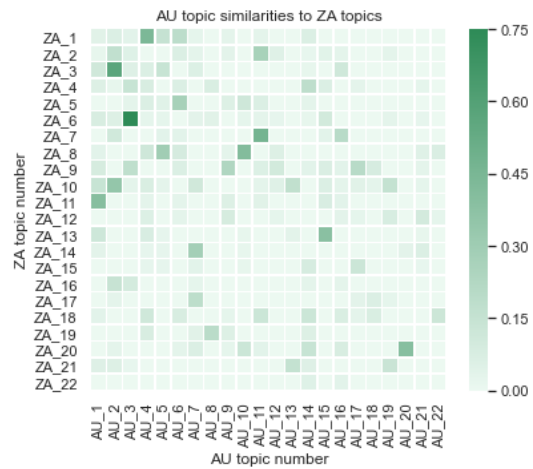


Figure 28: Topic cosine similarity between AU and ZA